2012

A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design

Erik Voss
Iowa State University

A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design

by

Erik Voss

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Major Professor
Dan Douglas
Volker Hegelheimer
Geoffrey Sauer
Amy G. Froelich

Iowa State University
Ames, Iowa
2012

Table of Contents

List of Tables vi
List of Figures viii
Acknowledgements x
Abstract xi
Chapter 1 Introduction 1
Chapter 2 Literature review 8
2.1 Purpose of collocation tests 8
2.2 Identification of a Target Language Domain 12
2.3 Collocation Selection 15
2.3.1 Word-based collocation identification 16
2.3.2 Statistical approach to collocation identification 23
2.3.3 Phrase-based collocation identification 24
2.4 Elicitation methods used in the test items 26
2.5 Scoring 32
2.6 Conceptual framework 35
2.6.1 Construct definition 35
2.6.2 Context of academic discourse 38
2.6.3 Metacognitive strategies 41
2.6.4 Relationships with other constructs in nomological network 43
2.7 Response formats 47
2.8 The interpretive argument 48
2.9 Research questions 56
2.10 Chapter summary 57
Chapter 3 Test development 58
3.1 Development relevant to domain description 58
3.1.1 TLU Corpus Selection and Description 58
3.1.2 Identification and selection of target collocations 63

3.1.3 Task Design 67
3.2 Development relevant to the evaluation inference 69
3.3 Development relevant to generalization inference 71
3.4 Development relevant to the explanation inference 72
3.5 Chapter summary 80
Chapter 4 Methodology 81
4.1 Design 81
4.2 Participants 82
4.3 Materials and instruments 84
4.3.1 Collocational Ability Test 85
4.3.2 Reading test 86
4.3.3 Vocabulary size test 86
3.4 Test Reflection Survey 88
4.3.5 Screen capturing observations 89
4.3.6 Questions for semi-structured post-interview 90
4.4 Procedure 90
4.4.1 Phase 1: Quantitative and embedded qualitative data collection 91
4.4.2 Phase 1: Quantitative data analysis 93
4.4.3 Identification of participants for qualitative data collection 98
4.4.4 Phase 2: Qualitative data collection and analysis 99
4.5 Chapter summary 100
Chapter 5. Results and discussion 101
5.1 Descriptive statistics and reliability estimates 101
5.1.1 Descriptive statistics and reliability: dichotomous method 102
5.1.2 Descriptive statistics and reliability: Polytomous method 104
5.2 Rasch IRT model fit 106
5.3 Reliability 119
5.4 Item facility and rank order 120
5.4.1 Correlation between rank order and item facility with dichotomous scoring 122
5.4.2 Correlation between item frequency and item facility with polytomous scoring. 124

5.5 Nomological network 125
5.5.1 Descriptive statistics for reading test 126
5.5.2 Correlations with reading test 127
5.5.3 Descriptive statistics for productive vocabulary size test 129
5.5.4 Correlations with vocabulary size test 130
5.6 Strategies and perceptions 133
5.6.1 Screen capturing data analysis 134
5.6.2 Analysis of Test Reflection Survey questions 1 and 2 137
5.6.3 Test Reflection Survey data: Question 3 140
5.6.4 Post-test interviews 146
5.7 Group differences 156
5.7.1 Distinctions among proficiency level groups using dichotomous scoring scale 157
5.7.2 Distinctions among proficiency level groups using polytomous scoring 158
5.8 Academic language 159
5.9 Student placement 161
5.10 Chapter summary 166
Chapter 6 Conclusion 168
6.1 Validity argument 169
6.1.2 Evaluation inference 175
6.1.3 Generalization inference 179
6.1.4 Explanation inference 181
6.1.5 Extrapolation inference 187
6.1.6 Utilization inference 190
6.1.7 Impact intention inference 191
6.2 Summary of validity argument 192
6.3 Limitations and implications 192
6.4 Suggestions for future research 195
6.5 Conclusion 196
References 197
Appendix A: Summary of task characteristics and variables 209

Appendix B: Test items 210

Appendix C: Target verb-noun word pairs sampled from BNC academic sub-corpus with most frequent collocate form 213

Appendix D: Concordance lines for verb-noun pair "spent ~ time" 214

Appendix E: COCA results and test-taker responses for "make ~ distinction".....215

Appendix F: Item-total statistics for dichotomous and polytomous data 217

List of Tables

Table 2.1. Summary of inferences, warrants and assumptions for the interpretive argument 55
Table 3.1. Raw joint frequency counts for constituency variations for the collocation “make ~ decision” 65
Table 3.2. Summary both scoring methods 70
Table 3.3. Total number of tokens and types for all test items (k = 35) 73
Table 3.4. Total number of tokens and types for all word pairs, collocates, and nodes (k = 35) 73
Table 4.1. Participants in the study by group 82
Table 4.2. Lists of target items on the vocabulary size sub-tests 88
Table 4.3. Test-taker responses and credit for the item for “make ~ distinction” 95
Table 4.4. Test-takers who participated in the post-interview 99
Table 5.1. Descriptive statistics for the Collocational Ability Test for all levels (k=35) using a dichotomous scoring method 102
Table 5.2. Descriptive statistics for the Collocational Ability Test for all levels (k=35) using a polytomous scoring method 104
Table 5.3. Standardized Rasch estimates for both scoring methods 107
Table 5.4. Item response theory (IRT) fit statistics for both scoring methods 110
Table 5.5. Rasch person separation and reliability estimates for both scoring methods 119
Table 5.6. Rank order of collocations by frequency in the corpus of written academic English 121
Table 5.7. Descriptive statistics for the reading test for all levels (k=35) 126
Table 5.8. Descriptive statistics for all three groups on Productive Vocabulary Size Test (k=30) 130
Table 5.9. Participants involved in screen capturing data collection 134
Table 5.10. Interviewees who participated in the post-interview 147
Table 5.11. Comments to the question about the purpose of the test 148
Table 5.12. Comments to the question about academic language 149

Table 5.13. Comments to the question about reading difficulty.....150
Table 5.14. Comments to the question about learning collocations 151
Table 5.15. Comments about Item 26 “pay attention”.....152
Table 5.16. Comments about Item 28 “receive attention”.....154
Table 5.17. Correlations for scores of both scoring methods on the Collocational Ability Test and subsections of vocabulary size test for all test-takers (N=206).....160
Table 5.18. Placement of test-takers by cut-scores.....164
Table 6.1. Warrants, assumptions, and types of backing for the domain description inference.....172
Table 6.2. Warrant, assumptions, and types of backing for the evaluation inference.....176
Table 6.3. Summary of support by scoring method for evaluation inference.....179
Table 6.4. Warrant, assumptions, and types of backing for the generalization inference.....180
Table 6.5. Warrant, assumptions, and types of backing for the explanation inference.....182
Table 6.6. Warrant, assumptions, and types of backing for the extrapolation inference.....188
Table 6.7. Warrant, assumptions, and types of backing or the utilization inference.....190
Table 6.8. Warrant, assumptions, and types of backing for the impact intention inference.....191

List of Figures

Figure 2.1. Howarth's (1998) collocation continuum....37
Figure 2.2. Theoretical relationships in nomological net....46
Figure 2.3. Interpretive argument structure for the Collocational Ability Test....50
Figure 4.1. Screenshot of the Collocational Ability Test....85
Figure 4.2. Screenshot of the productive vocabulary test....87
Figure 4.3. Overview of the mixed-methods research design used in this study....91
Figure 5.1. Histogram showing the distribution of scores on collocation test (k=35) or all groups (N=206)....103
Figure 5.2. Histogram showing the distribution of scores on collocation test (k=35) for all groups (N=206)....105
Figure 5.3a. ICC for item 30 (b=2.9)....111
Figure 5.3b. ICC for item 10 (b=-3.3)....111
Figure 5.4a. ICC for item 20 (b=-2.2)....113
Figure 5.4b. ICC for item 31 (b=-3.4)....113
Figure 5.4c. ICC for item 25 (b=-2.3)....114
Figure 5.4d. ICC for item 28 (b=-2.3)....114
Figure 5.4e. ICC for item 1 (b=3.5)....114
Figure 5.4f. ICC for item 2 (b=3.4)....115
Figure 5.4g. ICC for item 5 (b=2.6)....116
Figure 5.4h. ICC for item 7 (b=2.2)....117
Figure 5.5. Scatterplot showing the relationship between rank order for collocation frequency and item facility using the dichotomous scoring method (r_s = .02)....123
Figure 5.6. Scatterplot showing the relationship between rank order for collocation frequency and item facility using the polytomous scoring method (r_s = .40)....124
Figure 5.7. Histogram showing distribution of scores on reading size test by all groups (k=20)....127
Figure 5.8. Scatterplot showing relationship between dichotomous collocation test scores and reading test scores (r_s = .70)....128
Figure 5.9. Scatterplot showing relationship between polytomous collocation test scores and reading test scores (r_s = .62)....129

Figure 5.10. Histogram showing distribution of scores on vocabulary size test by all groups (k=30) 130
Figure 5.11. Scatterplot showing the relationship between Collocational Ability Test scores using dichotomous scoring and vocabulary size test total scores (r_s = .74) 131
Figure 5.12. Scatterplot showing the relationship between Collocational Ability Test scores using polytomous scoring and vocabulary size test total scores (r_s = .72 132
Figure 5.13. Average response change in percent by group (n = 8) 135
Figure 5.14. Responses about thinking in academic English while taking the test 138
Figure 5.15. Responses about language on the test and in university textbooks 139
Figure 5.16. Results indicating an ability to compare the texts 141
Figure 5.17. Percentage of distinctions indicating similar, different or both for each group 142
Figure 5.18. Box and whisker diagram showing score distributions for three proficiency levels on the Collocational Ability Test 157
Figure 5.19. Box and whisker diagram showing score distributions for three proficiency levels on the Collocational Ability Test using the polytomous scale 158
Figure 5.20. Overlapping of the three ability level groups' distributions on the Collocational Ability Test with trendlines 163
Figure 6.1. Backing collected for each step in the validity argument for the Collocational Ability Test 171

Acknowledgements

I am very grateful and fortunate to have had the opportunity to work with such wonderful people at Iowa State University. Each member on my dissertation committee has contributed to my growth in Ames both personally and academically. I am honored to have worked with my Major Professor, Dr. Carol Chapelle. Thank you for your guidance and support on research projects as well as the development and writing of my dissertation. I am sincerely thankful to Dr. Dan Douglas for his friendship and insights into the world of language testing. Your support and ideas have been inspirational. Thank you to Dr. Volker Hegelheimer for your encouragement from day one. To Dr. Geoff Sauer, I greatly appreciate your attention to detail and support through the Studio for New Media. I could not have accomplished the initial stages of test development as efficiently without the technological resources in the "Studio." I am very appreciative of Dr. Amy Froelich's guidance regarding my quantitative analyses and methodology.

Many thanks to those of you who contributed throughout the test development and data collection process: Dr. David Oakey for his shared interest in the probability of co-occurrences, Dr. Barbara Schwarte, director of IEOP, and the instructors at Iowa State University. I would also like to thank my family and friends for their patience and encouragement.

Abstract

Among the language capacities that L2 learners of English develop as they increase their proficiency is the ability to use appropriate collocations in the relevant registers of language use. Assessment of this capacity may therefore provide an efficient means of distinguishing among examinees with various levels of language proficiency within a particular register, although research has not yet attempted to operationalize such a measure. The purpose of this research is to explore the use of a measure of developmental, register-specific collocational knowledge as a means of making relevant distinctions among examinees in a minimal amount of time. Ultimately, such a test could be used in conjunction with other information for norm-referenced decisions such as placement, with a convincing argument for such uses. The first step, however, is to develop the interpretive argument and validity argument for score meaning of such a measure.

This validation study followed a mixed-methods embedded and sequential explanatory design consisting of quantitative and qualitative data collection (Creswell & Plano Clark, 2007). Quantitative data were collected from performance on the test of collocational ability, a productive vocabulary size test, and a reading test by 206 Chinese-speaking learners of English at three levels of English proficiency. Qualitative data were collected for a sample of the participants during and after the quantitative data collection and analyzed in order to identify evidence to explain initial quantitative results from the collocational ability test including test-taking processes. Qualitative data included screen capturing during test administration (n = 10), post-test interviews (n = 6) and a test reflection survey (n = 206). The test development, piloting, data collection, and analysis provide backing for the assumptions underlying inferences in the interpretive argument.

The study presents the development of a validity argument for the meaning of the score on the computer-based ESL collocational ability test, which could be used to contribute to the decision to allow a test-taker to participate in English-medium instruction or place the test-taker in an appropriate English language course. The first stage in the validity argument begins with the interpretive argument, which defines the grounds, inferences, warrants, and claims that provide the foundation for score interpretation (Kane, 2006). The claim that the interpretive argument needs to support is that the test score reflects the ability of the test-taker to use and understand lexical collocations as they are written in college and university settings. Backing for the inferences in the interpretive argument includes a theoretical analysis and empirical data. Theoretical evidence provides the backing for the domain definition, evaluation, and part of the explanation inferences. These empirical results provide evidentiary support for assumptions backing part of generalization, explanation, and extrapolation inference in the interpretive argument. Backing for the utilization and impact intention inferences is beyond the scope of this study. The backing for the assumptions that underlie the interpretive argument provide the basis for the validity argument.

Chapter 1 Introduction

Each year, a large number of students from non-English speaking countries enter college or a university in an English-speaking country such as the United States. The procedure for entrance often consists of one or more standardized language tests to assess their English proficiency. Admission to the institution is partially based on the score received from these tests. For those students who are admitted based on test scores and other materials submitted to the institution, further testing often takes place. Additional testing is needed either to allow students to matriculate directly into credit-bearing courses or place students into courses for supplemental English instruction, which focus on academic language and skills needed at the institution. These language tests at each institution place students into appropriate courses with the intention to assist with language instruction so the student can be successful in his or her studies. Such placement tests are usually administered in August or January at the beginning of a 14- to 16-week educational semester, and placement decisions are made before courses begin each academic semester. Students are placed according to the interpretation of the test scores on the placement test. Performance on the test is interpreted as an indication of a student's ability to comprehend and use academic English at an acceptable proficiency level. Scores from such tests are then interpreted as indicators of students' academic English ability and provide the basis for decisions about placement. Such inferences would ideally be made on the basis of samples of performance on academic tasks.

Whereas these placement tests intend to measure a student's language ability in an academic context, they are different from other language tests developed for another specific purpose or to measure general language proficiency. Assuming academic placement exams

are developed and used for their specific intended purpose, the development of these tests should be different from other tests. Douglas (2000) has identified two aspects that distinguish language for specific purposes (LSP) testing from general purpose testing, (a) authenticity of task, and (b) interaction between language knowledge and specific purpose content knowledge. These two criteria can be used to guide the development of academic placement tests.

The first criterion suggested by Douglas, authenticity of task, makes the connection between critical features of tasks in the target language use (TLU) situation of interest and the tasks on the test. In other words, the test tasks should represent the language that is used in the academic setting once the student is taking courses. At the beginning of the academic semester, however, the test battery is administered to a large number of students in such a short period of time it is difficult to replicate on a test the potentially complex real life tasks that are required of students at an institution of higher education (Douglas, 2010). Instead, test tasks need to incorporate essential language features so that the interpretations of performance are meaningful for the purpose of the test under the time constraints and available resources.

Placement exams generally measure reading, writing, listening, and possibly grammar. Measures of vocabulary are sometimes part of a scoring rubric, if measured at all. Most current placement tests do not have a direct measure of vocabulary. Scores on a reading test are used to interpret test-takers' level of reading comprehension. A writing assessment may be based on a holistic rating scale, which most likely involves a subjective evaluation of the appropriateness of vocabulary used in the writing sample. Furthermore, scoring rubrics generally account for vocabulary at the single-word level. Other dimensions of vocabulary

are not taken into consideration. The two commonly discussed dimensions of vocabulary are breadth and organization (depth). *Breadth* refers to the size of a vocabulary or how many words a student knows. *Organization* or depth of vocabulary describes how well the words are known, the different meanings of a single word, or knowledge of other words that frequently co-occur when produced. The ability to produce a combination of words appropriately is also called *phraseological* or *collocational* ability. Pawley and Syder (1983) suggest that collocations may be involved in native-like selection as well as native-like fluency and are becoming more relevant in language assessment.

Collocations are generally grouped into two types: grammatical and lexical (Benson et al., 1986). *Grammatical collocations* are phrases with a dominant word, noun, adjective, or verb, followed by a preposition, infinitive, or clause. *Lexical collocations* are usually constructed with nouns, adjectives, verbs, and adjectives. Grammatical collocations collocate a dominate word with subordinate word such as a preposition, whereas lexical collocations are comprised of two words "equal words" (Schmitt, 2000). For instance, *agonize over something* is a grammatical collocation containing a preposition, whereas *make a decision* and *weak tea* are examples of lexical collocations. Because common verb-noun lexical collocations have been found to cause difficulty for second language learners (Bahns & Eldaw, 1993; Biskup, 1992; Howarth, 1996, 1998; Nesselhauf, 2003), these collocations were the target collocation type in this study. Verb-noun collocations are further defined as restricted collocations. Collocations are considered restricted in the number of verb combinations that are possible with a particular noun. Previous studies have also found the use of restricted collocations difficult for L2 learners. Considering the difficulty language learners have with restricted verb-noun collocations, a measure of this language feature

might help distinguish students who are ready for independent academic study from those who need additional instruction.

The addition of a measure of collocational ability developed using restricted collocations may provide supplementary information about language ability, which is intended to reflect language use in courses at an English-medium colleges and universities. It is possible that a test of collocational ability could be integrated into an existing English placement test as an additional measure of phraseological ability that may not be accurately measured by current placement tests.

Douglas's (2000) second criterion for distinguishing language for specific purposes (LSP) testing from general-purpose testing is interaction between language knowledge and specific-purpose content knowledge. Background knowledge can affect performance on an assessment for test-takers who are familiar with the content and/or context of the test. It can also make a test more difficult for test-takers with very little knowledge of the test content or context; therefore, test-takers' background knowledge may contribute to error in measurement.

In the test development process, once the purpose of the test had been established, a description of the target language domain and task is needed as a way of controlling for specific purpose content. The language features can be sampled from the target domain and operationalized in a test task. Because language use is embedded in particular situations, and each instance is unique, it is impossible to list all of the possible instances (Bachman & Palmer, 1996); however, corpus tools allow us to sample language and view a large number of instances in particular situations. Sampling collocations from a corpus of the target language domain can be useful in developing test items that connect language knowledge and

specific purpose content knowledge. Large corpora are useful for identifying language features by frequency offering advantages to sampling language in this manner. One reason for selecting high-frequency collocations is what Laufer and Nation (1999) call a *cost-benefit distinction*. The time and effort to learn a lexical item is best spent on items that will be used frequently. "It is a commonly accepted truth in foreign language learning that the more frequent a word is in a language then the more easily, and the earlier, it is likely to be learned" (Milton, 2007, p. 48). This concept is also congruent with the interactionalist perspective by identifying relevant language features in discourse; thus, if specific-purpose content knowledge is part of the construct to be measured, corpora can be a resource for sampling relevant and representative language features in the target language domain.

The primary purpose of a language test is to allow inferences to be made about learners' language abilities (Douglas, 2010, p. 17). The test of collocational ability in this study applies Douglas's criteria for designing a language test for the specific purpose of eliciting performance of collocational ability as a language feature in the context of academic English in order to provide a meaningful score that can be interpreted to contribute to the decision to allow a test-taker to participate in English-medium college or university or place the test-taker in an appropriate English language course.

The test of collocational ability is then evaluated using an argument-based approach to validation. Each step in the argument is intended to support the interpretations, uses, and impacts of the test scores. The claim that the interpretive argument needs to support is that the test score reflects the ability of the test-taker to understand and use restricted verb-noun collocations as they are written in college and university settings. The argument-based approach begins with an interpretive argument (Kane, 2006). The interpretive argument

outlines the grounds, claims, warrants, and assumptions for each inference in the argument (Chapelle, 2008; Chapelle, Enright, & Jamieson, 2010). The approach consists of a chain of inferences each based on a unique warrant, which is supported by assumptions. The final inferences in the argument lead to claims about test use and intentional impact of the score interpretation and score use. This approach to validation follows practical logic by seeking theoretical and empirical evidence to support each inference. Adequate backing for the assumptions underlying each inference is required before moving to the next inference in the chain. The evidence collected for the interpretive argument is then evaluated and presented as a validity argument. Although the final goal is to present a validity argument for a test based on the scores from the tasks, the interpretive argument should also be used as a design plan at the beginning of the test development process (Briggs, 2004). Formulation of an interpretive argument at the beginning of the test development process can guide the development of the test so that the interpretation of test scores is meaningful for the intended uses and impacts that are desired.

The knowledge acquired in this study has the potential to make contributions to the study of collocations and validation of language assessment. One benefit of this study is the contribution it can make to approaching and defining a construct of collocational ability. Many studies have measured collocational ability through various sampling methods, making it difficult to compare results. Ideally, the findings from this study can contribute to refining a framework and construct for measuring collocational ability.

Secondly, this study is intended to identify collocational ability as a feature of language that might be useful as part of placement decisions at English-medium colleges and universities. The backing and challenges presented in the validity argument can lead to

better-informed test development of more sensitive measurement instruments for making such placement decisions.

A third benefit of this study is the awareness raised regarding the importance of introducing collocations in English language instruction. A greater awareness of the importance of collocations in English language instruction and language learning materials offers the potential for intended consequences in the form of positive washback.

Finally, this study contributes to the awareness of partial credit scoring for short constructed response tasks. Dichotomous scoring methods for such tasks do not award any credit for responses that demonstrate partial knowledge. Scoring procedures should be developed so that test-takers receive credit for what they know.

The following chapter reviews the literature on assessment of collocational knowledge, describes the framework for measuring collocational ability, and introduces the interpretive argument. Based on the interpretive argument, seven research questions are identified that will be addressed through analysis of empirical data. The third chapter details the test development process for the Collocational Ability Test. Chapter 4 describes the methodology for the study including the overall mixed-methods research design, participants, materials, data collection, and analysis. Chapter 5 presents and discusses the results of the study as they pertain to the research questions. The final chapter, chapter 6, concludes with a summary of the dissertation and a presentation of the validity argument with support from the theoretical and empirical evidence presented in the previous chapters. The chapter concludes with implications and limitations of the study.

Chapter 2 Literature review

This chapter outlines the concepts that underlie and inform the construct of collocational ability and the assessment instrument for this study. The chapter begins with an outline of former and current perspectives on measuring collocational knowledge including test purpose, item selection, task types, and scoring methods. The interactionalist approach is then introduced as the underlying conceptual framework for language ability construct that guided test development and validation in the study. Next, the interpretive argument for test score interpretation is presented, detailing the warrants and assumptions that require backing to support a validity argument. The chapter ends with the research questions for the study. The development of the research questions was guided by identifying which evidence was needed to provide backing for the interpretive argument. The questions focused on the weaker assumptions that would benefit from evidence to either support or challenge the warrants in the interpretive argument.

In the last two decades, only a few studies on the assessment of collocational knowledge by L2 learners of English have attempted to address the issues involved in the development of measurement instruments. The following review will highlight issues related to test purpose, identification of a target language domain, diverging approaches to the selection of frequent collocations in target language corpora, and the type of elicitation task.

2.1 Purpose of collocation tests

Tests used for making inferences about collocational knowledge typically have not contributed to educational admission or placement decisions but instead appear in research intending to develop experimental item formats and explore the construct of collocational

knowledge. Studies in the 1980s and 1990s were designed to raise the important issue that L2 learners lack collocational knowledge. Such early measures of collocational knowledge aimed at exploring whether collocations need to be part of explicit instruction. This research area was characterized by a focus on classroom participants from a single proficiency level, which was usually advanced. Moreover, each study was conducted with a sample of English as a Foreign Language (EFL) learners comprised of a single L1, e.g., Arabic (Farghal & Obiedat, 1995), German (Bahns & Eldaw, 1993; Channell, 1981; Herbst, 1996), and Greek (Gitsaki, 1999). Only a couple of studies have included multiple L1s in an English as a Second Language (ESL) context (Aghbar, 1990; Bonk, 2001).

Early research began in the classroom. Channell (1981), for instance, asked a class of eight German-speaking learners of English with advanced proficiency level to fill out a grid indicating which adjectives collocate with a list of nouns. Channell's interpretation of the performance on this task indicated that L2 learners did not realize all of the potential combinations even though they knew all individual words in the grid well. Productive responses were also elicited with the purpose of measuring general knowledge of collocations by L2 learners of English. In a study by Bahns and Eldaw (1993) using a translation and a cloze format, German advanced EFL students were asked to produce 40 collocations. Similar to Channell's (1981) findings, Bahns and Eldaw (1993) concluded that the production of collocations for correct English presented a problem for advanced learners. The practical use of these early measures was used to inform classroom practices.

In the early 1990s, the purpose of elicitation measures extended from determining that collocations were difficult for L2 learners to discovering how collocational knowledge by L2 learners compares with that of native speakers. Aghbar (1990) conducted the first

study that included participants with multiple L1s. His purpose was to compare the production of verb-noun collocation pairs in 50 sentences by 97 ESL Freshmen, 44 target language university students, and 27 university faculty. The faculty produced the greatest number of appropriate collocations, followed by the American students, and then the ESL students, demonstrating a distinction between native speakers (NS) and non-native speakers (NNSs). Research continued with comparing L2 learners' knowledge of collocations with target language speakers' knowledge. On a larger scale, Herbst (1996) included 100 students from two German universities and 58 university students in England in translation and cloze tasks. The collocational production of the English speakers was more "uniform" than that of the responses by the German students.

A similar study by Farghal and Obiedat (1995) compared the collocational knowledge of 34 junior and senior university-level Arabic-speaking English majors and 23 Arabic-speaking teachers of English who were considered high-proficiency participants. The results indicated that both groups were deficient in their knowledge of English collocations. The teachers of English in this study were more similar to advanced learners than target language (NS) test-takers. Accordingly, findings from early research comparing collocational knowledge by NS and NNS indicate that recognition and production of collocations by L2 learners was not the same as production by target language speakers in either quantity or quality.

Through the nineties, research proceeded with homogenous L2 learners classified either as having an advanced level or university-level language proficiency, which many consider as an advanced level of English. Results from comparison studies with target language users indicated that L2 learners were deficient in their knowledge and production of

collocations based on the interpretations of the test scores. The research agenda expanded near the end of the decade. Research on collocational knowledge started to include participants from a variety of proficiency levels for groups with single and multiple L1s.

Following the new agenda, two studies comparing collocational knowledge with a measure of general proficiency included participants with single and multiple L1s (Bonk, 2001; Gitsaki, 1999). In the study by Gitsaki (1999), 275 Greek students in the first, second, and third years of junior high wrote essays, performed translation tasks, and completed a cloze task. The study was comprehensive including grammatical and lexical collocations in addition to a number of different subtypes of collocations, but was limited to Greek-speaking learners of English. Results indicated that lexical collocations are learned after grammatical collocations. Gitsaki did not report reliability estimates and item statistics. Bonk's (2001) participants made up a wider span of proficiency levels, from low-intermediate to advanced. These 98 participants completed a number of tasks including a 20-item cloze instrument measuring knowledge of verb-noun collocations. Bonk's (2001) study, although smaller in scope, supported the findings by Gitsaki (1999) with a more diverse range of L1 among the participants, although with fewer collocation subtypes. Both studies found a positive relationship between knowledge of lexical collocations and proficiency level. Bonk's (2001) study reported a correlation coefficient of $r = .73$ after correction for attenuation, significant at $p<.05$, 1-tailed test.

The research purpose of most tests in the 1980s and 1990s resulted in participants from a single proficiency level and a single L1 being chosen as test-takers. Past tests of collocational knowledge have served solely as a tool to make inferences about the receptive and/or productive knowledge of collocations by L2 learners of English and to investigate the

difference in collocational knowledge between L2 learners and target language speakers. Results from these studies suggested the need to teach collocations in the classroom, revealed the degree of collocational knowledge by L2 learners in relation to general language proficiency, and provided response comparisons between native speakers and non-native speakers. No study to date has included a measure of collocational ability as part of a decision about placement of language learners in appropriate levels of English instruction. Since the purpose of my study is to investigate the potential for scores from a test of collocational ability to contribute to and strengthen placement decisions at an English-medium college or university, test-takers were recruited from a variety of proficiency levels to allow for interpretations to be made regarding collocational ability at the various levels. In addition, the use of test-takers with a single L1, Chinese, was intended to keep the first language variable constant and L1 transfer fairly consistent for the participants in this study. L1 transfer is addressed below.

2.2 Identification of a Target Language Domain

In the past 30 years, researchers have developed measures of collocational knowledge, and the selection of collocations on the tests has been at the discretion of the test developer; thus, the domain from which the collocations were selected was either specific to instruction or from unrelated lists or collections. Collocations were selected from textbooks, dictionaries, collections of idiomatic English, and intuition. Channell (1981), for example, sampled items from a book called “The Words You Need” (Rudzka, Channell, Putseys, & Ostyn, 1981) for her collocational grid. This book is a textbook co-authored by Channell herself and includes definitions and sentence examples for “tough-but-common words.”

Bahns and Eldaw (1993) used the book co-authored by Channell as well as two books that focused on idioms and collocations respectively, *English Idioms and How to Use Them* (Seidl & McMordie, 1978) and *The BBI Combinatory Dictionary of English* (Benson, Benson, & Ilson, 1986), to select items for a translation task and a cloze-type task. These resources provided collections of collocations for both classroom instruction and test development.

Other studies have failed to indicate the source of their item selection. Aghbar (1990) did not state explicitly where the items for his cloze test came from but provided examples from the BBI dictionary (Benson et al., 1986). Herbst (1996) did not indicate how his items were selected either but mentioned the following dictionaries: the *Oxford Advanced Learner's Dictionary* (OALD) (Crowther, 1995), the *Longman Dictionary of Contemporary English* (*LDOC*) (Summers, 1995), and *Collins COBUILD English Dictionary* (Sinclair, 1995) and the use of test items that are identical to those of Greenbaum (1988). Both Gitsaki (1999) and Bonk (2001) reported developing their tests based on collocation types found in the BBI dictionary (Benson et al., 1986).

Although the source of collocations is fairly general and ambiguous in previous studies, an important observation from Aghbar's (1990) study found that less frequent or more formal items from written English on a test would most likely show greater difference between target language users and L2 learners. This observation led to the suggestion that the proper use of collocations in their proper register is "an important aspect of language ability" (p. 7). This recommendation is in line with Sinclair's (1966) observation that collocations are best studied in a particular register or situation of use. Aghbar's (1990) recommendation to identify collocations by register is congruent with the concept of identifying a target

language use domain in language testing. Whereas language use is embedded in particular situations, and each instance is unique, it is impossible to list all of the possible instances (Bachman & Palmer, 1996); therefore, collocations can and should be sampled from a corpus of the target language use domain.

As indicated above, one characteristic of most previous studies is the item selection process from an unknown domain with little information given regarding where collocations were selected for each study. Additionally, even less information is available regarding how the collocations were selected. One study, for example, that included collocations from common everyday topics associated with food, clothing, and the weather did not explain where they came from or how the items were selected (Farghal & Obiedat, 1995). Although the collocations were still sampled from "general English" topics rather from a specific genre or register, this study was a step closer to identifying items in a particular target language situation.

The selection of items from a specific target language domain such as an academic language domain was something that few studies had attempted before. One experimental study isolated target nouns from an academic English domain by selecting nouns from the 1995 Test of English as a Foreign Language (TOEFL) for a sentence elicitation task (Schmitt, 1998). The nouns were used as prompts from which test-takers were asked to write sentences, which would then be analyzed for use of acceptable collocations. Unfortunately, a number of participants did not know the meaning of a few of the single-word prompts or nodes, which made the measurement of collocational knowledge difficult. Consequently, collocational knowledge can be difficult, if not impossible, to measure if one or both of the individual words in the collocation are not known by the test-taker. Nevertheless, Schmitt's

methodology for test construction is the closest approach to the selection process used in my study. This difference, however, is that Schmitt sampled his collocations from a TOEFL test, whereas my study utilized collocations as a linguistic feature of authentic academic language for a test of collocational ability.

In light of the factors identified that affect performance on a test of collocational ability, my research selected items from general academic discourse with the purpose of developing a test that would contribute to placement decisions. The collocations for the Collocational Ability Test in my study were sampled from a specific register of written academic discourse. A range of disciples is necessary to avoid test favoring test-takers with specific topic knowledge. The files in the academic section of the BNC from which the collocations were sampled consist of a number of academic disciplines, including technical engineering, social science, politics and law, natural science, medicine, and humanities and arts. These files represent academic English referenced in the rest of this paper as general academic discourse or general academic English.

2.3 Collocation Selection

The next section will describe approaches taken with target language corpora that address the issue of how items are selected. The frequency of individual items in the composition of the collocations and their influence on the knowledge and use of the collocation are discussed.

2.3.1 Word-based collocation identification

Issues addressed thus far have been related to test purpose, target language domain, and source of collocations. The construct-based inference of these tests was collocational knowledge or ability; however, the use of the tests was less clear. Items selected for the collocation tests originated from intuition or reference materials without specifying a particular context or use situation. In addition, the tests did not include a systematic selection of items that could be used as a diagnostic test nor were they from a particular target language use domain that could be used to make decisions about examinees' performance in a particular non-test situation. The vast number of collocations adds to the difficulty of selecting representative items in general English.

In the early 2000s, this issue of how to select collocations was addressed in a number of studies. The key concept in item selection at this time was *frequency*. The concept stems partly from the definition of a collocation as words that frequently co-occur. Frequency-based selection rests partly on the idea that collocational knowledge cannot be measured if one of the words in the collocation is not known and partly on the assumption that words with high-frequency should be prioritized, because they are most likely to be encountered or used in everyday language use since they are "frequent." This frequency-based approach was facilitated, if not inspired, by the emergence of corpus linguistics.

The use of target language corpora as a resource for test development is not unusual. A corpus analysis has been used to identify vocabulary items for language testing for high-stakes tests as well as to inform item development for reading and listening for the TOEFL (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Enright et al., 2000). More specifically,

to base their testing materials on real texts, Cambridge ESOL selected collocations using English target language corpora (Barker, 2004).

A systematic selection of items based on frequency from a large target language corpus holds promise as an effective method for identifying relevant collocations for language test development. As English language corpora became available, a number of studies extracted collocations for test item development using corpus-based techniques (Gyllstad, 2005; Mochizuki, 2002; Moreno Jaén, 2008; Revier, 2009). Studies sampling from corpora in the early 2000s were investigating the knowledge of collocations by L2 learners based on highly frequent collocations.

In the first decade of this century, a method for using native speaker corpora to identify target collocations became popular through a word-based approach beginning with single-word frequency lists (Eyckmans, 2009; Gyllstad, 2005, 2009; Mochizuki, 2002; Moreno Jaén, 2007; Revier, 2009). From a word list of the most frequent words in a corpus, nodes were selected as either nouns or verbs, then collocates were identified for each target node. The selection process was described in more detail in most of these studies but not adequately enough to know how the items were selected from the word frequency list.

Mochizuki (2002), for example, designed a test of collocational knowledge beginning with a list of high-frequency verbs from a target language corpus and then identified collocates for the selected words. The collocation test was developed by identifying collocates for 18 words, six nouns, six verbs, and six adjectives, that had been chosen from four frequency lists, although his report does not mention how this was done. The collocates were then identified using Collins COBUILD English Collocations (1995) on CD-ROM and the interactive version of the Edinburgh Associative Thesaurus (EAT)

(http://www.eat.rl.ac.uk/). Once again, the sources were identified but the process of selecting an appropriate collocate was not explained. In some instances, a number of different verbs may be appropriate for a single noun in a different use situation. Theoretically, this method of item selection based on frequency of individual items should produce tests with items that are familiar to the test-takers. Results from this study indicate, however, that performance, although low on the collocation test, was still greater than on both the vocabulary size test and the test of paradigmatic knowledge, which were also constructed based on word frequency.

Similarly, Moreno Jaén's (2007) sampling approach began with 80 high-frequency words from a word frequency list based on the British National Corpus (BNC) the Bank of English and the Longman Corpus. Moreno Jaén (2007) took a multiple corpus approach sampling high-frequency items from three very large corpora with the intention of sampling from a larger domain claiming that this improved the validity "of the test." Collocates for these words from the list were then identified in the BNC using Wordsmith tools software. Even though a word list created from three corpora was used to identify nodes, the collocates were chosen from a word list comprised of collocation for a single corpus. Since collocation pairs are not as frequent as individual lexical items, it might be possible that the frequency of a collocation pair in one corpus is quite different from the frequency in a different target language corpus. The use of multiple corpora may have confused the selection processes rather than improving it. Performance on both the receptive and productive measures in this study indicated that university-level English majors in Spain had a low general knowledge of frequent collocations when they were selected in this manner.

Another study that began with a word list developed instruments to measure collocational knowledge by 25 high-intermediate level students in Belgium majoring in translation and interpretation after 60-hours of instruction (Eyckmans, 2009). Initially, 40 verbs were selected from a frequency list as nodes. Collocates for this test were then identified in the BNC using a probability statistic called the z-score. A z-score is a statistical calculation of the probability for two words co-occurring. A second corpus was used to identify the frequency of the collocation pair in the Collins COBUILD Bank of English. Test-takers were asked to distinguish between real and pseudo collocations that were developed for the judgment task. Performance on the post-test was higher than on the pre-test; however, the reliability estimate dropped from .90 to .64 from the first to the second. Less variance in the post-test, in this case, would be considered an indication of improvement.

Beginning with a list of high-frequency verbs taken from the Japan Association of College English Teachers word list known at the JACET 8000 word frequency list (Iwashita, et al., 2003) created using the BNC, Gyllstad (2005, 2009) used a z-score to identify noun collocates in the BNC. Gyllstad developed his testing instruments after sampling collocations from the BNC, selecting word combinations where both constituents were comprised of high-frequency words. Because all words were identified as high-frequency words, he hypothesized that there would be a great possibility that test-takers would know both words in the collocation. These collocates were used in receptive recognition formats, a judgment with real and pseudo-collocations and a matching task. Performance on these two test formats by 307 Swedish university English majors as a group was found to correlate positively with a measure of vocabulary size; however, performance on the collocation test

did not produce significant differences among all proficiency levels identified by year in school. The three proficiency levels were determined by level of education including students from first, second, and third term at the university. It is necessary to keep in mind that the independent frequency of each part of a collocation does not necessarily indicate knowledge of the two words together as a collocation.

In addition to selecting items by frequency, another study by Revier (2009) focused on exploring whether production of collocations is influenced by their semantic properties. The semantic properties were based on the degree of transparency of meaning of each collocation. The sampling approach for this study claimed to take a phrase-based approach by identifying collocational pairs based on the frequency of both words in a collocation. Rather than selecting single items as nodes, as in previous corpus-based research, collocations as multiword items or single units were selected from the BNC using the Phrases in English (PIE) interface (Fletcher, 2003); however, this was done by identifying adjacent words for 15 pre-selected verbs. These word pairs were selected if they were found to have a frequency range from .04 to .47 per occurrences per million. Revier was more explicit regarding the criteria for selecting the collocations that were identified by the Phrases in English interface. Performance on a 45-item collocation test by fifty-six Danish learners of English at three proficiency levels indicated significant differences between the high school groups and the more advanced university group. Revier (2009) reasoned that the frequency of the noun in the collocation was an important factor in the test results and suggested that low-frequency nouns should be replaced with high-frequency ones. Once again, the possibility is that individual words in a collocation may not be known by the test-taker, thus, making the measurement of collocations impossible if all of the constituents in the collocation are not

known. Potential avoidance of unknown words by the test-taker is another promising role that frequency may play in selecting items on a collocation test; however, opinions differ on the method of selecting collocations. If collocations are selected based on the frequency of the individual constituents, the frequency of the collocation is artificial and does not represent the actual frequency of the collocation.

Following the assumptions associated with the frequency of lexical items, Durrant (2009) explored the development of a list of academic collocations from a sub-corpus of the BNC. The BNC is a large corpus that has often been used to identify collocations for assessment instruments. Durrant's list contained very few collocations made up of combinations of verbs, nouns, adjectives, or adverbs, known as *lexical collocations*. The majority were *grammatical collocations*, comprised of combinations of a verb or adjective followed by a preposition, which are also very common words. Durrant warned that if frequency lists are limited to words that are found only in academic texts, a large number of combinations that are essential to academic discourse may be missing, even though they are also common in other genres or registers. This is relevant to the way in which collocations have been selected in previous research by sampling from a frequency list and then finding collocates in the same domain for the preselected target nodes. This approach does not identify the frequency of the collocation but instead identifies a collocation of unknown frequency based on the frequency of individual parts.

This aspect of selection has been consistent with every corpus-based study so far. Even the study by Revier (2009), which selected the collocations as whole units, began with a list of 15 verbs that needed to be included in each pair. This approach also restricts the sampling process to a list of high-frequency verbs (or nouns) and their collocates rather than

identifying collocations based on the co-occurrence of both the node and the collocates. This selection method has been known as the *word-based* approach, because the process begins with a list of single words.

Mochizuki (2002) and Gyllstad (2005, 2009) were careful to select high-frequency words for both constituents in the collocation pair; however, the words were selected individually, not together as collocation pairs. Gyllstad admitted that a limitation of the sampling procedure was the word-based approach in which single words were used as the nodes, and collocates were determined later. Unfortunately, not all researchers describe their selection process with enough detail to replicate the selection process. Moreover, if collocations take a meaning other than that of each of their constituents, the individual frequency of each word may not be as relevant as the frequency of the collocation itself in a specific context. As collocations are selected as units rather than individual constituents, an interpretation of a score can be that a test-taker does not have knowledge of a collocation as a single unit but rather only knowledge of individual lexical items.

Thus far, the word-based approach has dominated research that uses corpora to select word pairs for items to measure collocational knowledge, which is similar to the perspective taken in presenting collocations in dictionaries from a lexicographic perspective (Oakey, 2009). In this approach, the researcher focuses on a particular word of interest and searches for collocates of that particular word-form or set of inflected forms (Clear, 1993). This approach is also altered by early human intervention whereby the researcher places restrictions on the word combinations that are identified using corpus tools. A biased measurement is often the result of the frequency of a collocation, because it favors one of the two constituents in the collocation. This word-based approach has limitations and raises

questions as to the definition of a word as single or multiple units and the validity for the construct of *WORD* comprised of multiple lexical items (Gardner, 2007). An alternate method is the statistical approach, which can be used to identify multiword lexical items in a large corpus.

2.3.2 Statistical approach to collocation identification

The statistical approach employs statistical probability to identify collocational pairs or phrases. Word pairs can be identified on the strength of the co-occurrence of two words. This approach identifies word combinations using surface features of the items in the corpus and ignores any meaning based co-occurrences. A method for identification of word pairs using mutual information (MI) score and T-score statistics finds co-occurrences of two words based on probability of a word pair rather than the frequency of the individual constituents. An MI score measures the strength of the co-occurrence of two words and can be compared across corpora, whereas a T-score measures the certainty of two words co-occurring and is influenced by the size of the corpus (Hunston, 2002). MI- and T- scores have not been employed in the identification of collocations for test development. On the other hand, a z-score has been used to verify collocates in the BNC (Eyckmans, 2009; Gyllstad, 2005, 2009). Eyckmans (2009) used a z-score to verify verb-noun combinations in the BNC and Collins COBUILD Bank of English, which were initially identified by searching for collocates form a list of frequent verbs. Similarly, using a z-score, Gyllstad (2005, 2009) verified the strength of collocations after collocates had been identified for a list of highly-frequent verbs. Alternate approaches based identification and classification of word combinations on meaning as well as co-occurrence (e.g., the phrase-based approach).

2.3.3 Phrase-based collocation identification

A different perspective from the word-based approach includes the notion of co-selection of words, whereby speakers and writers select the word combinations together to create meaning (see Cheng, Greaves & Warren, 2006). A writer or speaker selects a combination of words based on the meaning of the utterance rather than selecting one word at a time. Selecting words together based on frequency was the method that Revier (2009) intended to use in the selection of target items for his test of "whole collocations". In the first phase of his selection process, using pre-selected verbs, Revier employed PIE to find verb-noun combinations in the entire BNC. Revier's method recognizes the need to address the issue of the shift in meaning by searching for co-occurrences of both words together rather than the combination based on the frequency of the individual words. Due to the co-occurrence of the words, Sinclair called this shift in semantic meaning a Meaning Shift Unit (MSU) (Cheng et al., 2009); however, although Revier (2009) identified word pairs based on the frequency of the words together, his process included early human intervention by identifying a list of verbs that were required to be part of the collocations. This is not unlike the word-based approach, wherein target nodes are pre-selected. Human intervention can be early or later in the process. In a process that favors identification of collocations by frequency, a later human intervention approach is desired. Human intervention is necessary, because computers cannot yet identify the semantic properties and/or shift in meaning of collocations that are necessary to distinguish collocations from free combinations of words. At this point, test developers do not rely solely on technology for test development. "UCLES continues to use the intuition, knowledge and experience of its item writers, examiners and subject officers in developing suitable tasks and topics for EFL examinations" (Ball, 2001, p.

6). Whether it is early or late, human intervention is used at some point in the identification process.

The identification of a collocation as a single unit including all of the words in the collocation is known as the *phrase-based* approach. The selection of collocations based on the frequency of both items in a collocation pair would identify its actual frequency as a single lexical unit in a corpus. One recent new analytical tool that holds promise for identifying collocations as phrases in target language corpora is ConcGram 1.0 (Greaves, 2009). ConcGram offers an alternative to the traditional corpus tools that search for contiguous n-grams. An n-gram is a series of consecutive lexical items. The "n" represents the number of constituents in the series. A series of two items, for example, would be a bi-gram. Likewise, a series of three items would be a tri-gram. The software offers a way to identify phraseological variation by identifying sets of words that co-occur despite their constituency variation or positional variation (Cheng, et al., 2006). Initially, Concgram uses raw frequency counts rather than probability algorithms such as the MI- and T-scores in the statistical approach to identify two or more word combinations in a predetermined span. In addition, ConcGram identifies word combinations regardless of constituency (contiguous and non-contiguous) or positional (node – collocate and collocate – node) variation.

This study used ConcGram 1.0 software to identify the raw frequency of co-occurring verb-noun word pairs to identify the frequency count of the verb-noun pairs rather than the frequency based on individual constituents in the pair. The software was used to search a large academic sub-corpus of the BNC (Lee, 2001) to identify verb-noun co-selections (MSUs) that could be used to create an assessment instrument to measure L2 learners' academic collocational ability. This test development process selected lexical collocations

from a target language corpus of academic English without eliminating words, as suggested by Durrant (2009), which are not exclusive to this register and used late human intervention to identify collocations based on frequency of the word combination first and then manually selected the items after they have been identified.

2.4 Elicitation methods used in the test items

After collocations have been identified systematically in the target language domain of interest, an appropriate task can be developed to elicit performance from which language ability can be inferred related to the purpose of the test. Language assessment tasks should elicit language characteristics that are representative of the more complex real-life tasks; yet, constraints on time and other resources may limit the extent to which test tasks can model the detail of complex real-life tasks. Previously, instruments measuring collocational knowledge have included receptive test formats, collocational grids (matching), acceptability judgment tasks, multiple-choice items, productive tasks, sentence elicitation, translation, and cloze-type items (Bahns et al., 1993; Biskup, 1992; Bonk, 2001; Farghal et al., 1995; Gitsaki, 1999; Gyllstad, 2005). Issues related to elicitation methods in previous studies on collocational knowledge are receptive vs. productive ability, contextualized vs. non-contextualized test items, and task characteristics that discourage the use of negative language strategies including avoidance, paraphrasing, and L1 transfer. These issues are discussed here within the context of two task categories: *selected* and *short response*. Because the focus of this paper is on developing a short response test format, the extended response format is not discussed here.

2.4.1 Selected response tasks

One group of studies developed receptive measures using selected response tasks such as collocation grids (Channell, 1981), multiple-choice items (Mochizuki, 2007), judgment tasks (Eyckmans, 2009; Gyllstad, 2005, 2009), and matching (Gyllstad, 2005, 2009). Using a non-contextualized item type, Channell (1981) remarked that learners often fail to mark correct collocations with recognition tasks that present multiple choices to the learner such as her collocation grid. It is difficult to know if the learner does not know the collocation or just did not notice the combination at that time. Moreover, Shillaw (2009) suggested that non-contextualized receptive judgment tasks may be measuring "something more like sensitivity to collocational potential, rather than actual knowledge of collocations" (p. 177).

Collocation grids, judgment tasks, and matching tasks often contain non-contextualized word combinations. An example of a contextualized selected response item is a sentence completion with multiple choice answers (Moreno Jaén, 2007). Although this item format is similar to a gap-filling item type, the potential responses for the gap are listed with the item. Performance on this receptive measure was higher than the gap-filling productive measure which was also included in Moreno Jaén's study. Revier (2009) produced a similar experimental task. His test-takers had to select a combination of node, collocate, and a potential article, if necessary, from three jumbled potential collocations. This test format was capable of discriminating among three categories of collocation transparency but not among proficiency levels.

Results on receptive selected response measures have indicated some knowledge of collocations but are prone to guessing by the test-taker. A test-taker may mark a collocation

even though it is not known as well as fail to select a collocation because he/she may be uncertain. In addition, the receptive measures do not indicate if a test-taker is able to use a collocation in context; thus, research on collocational knowledge using productive elicitation measures is more relevant to the focus of this study, which measures collocational ability.

2.4.2 Short response tasks

A short response task requires the test-taker to produce either a complete collocation or part of a collocation. Such productive measures (i.e., translation, sentence elicitation, and cloze item types) require learners to produce lexical items eliciting evidence of collocational knowledge with less possibility of guessing. In a free production task, test-takers often use strategies to produce language. These strategies may be useful in a communicative setting but are not always useful in a testing situation where a target collocation is elicited but not produced.

One type of short response format typical of the early studies was a translation task. Many early studies used translation of individual sentences containing a collocation as an elicitation instrument (Bahns et al., 1993; Biskup, 1992; Farghal et al., 1995). In a translation task, test-takers were asked to translate a sentence from their L1 that contained the equivalent of a target collocation in English. Unfortunately, the dynamicity of language allows translation and interpretation of a sentence in a variety of ways that could be considered grammatically correct yet may not contain a target response. The language produced in such translation tasks often did not contain a target collocation at all. Using strategies such as paraphrasing and avoidance, target collocations may be avoided, making it difficult to

analyze production of a target feature. As a result, translations resulted in learner production that was not relevant to the purpose of the test.

Strategies used in translation tasks vary by first language. For example, by comparing 34 Polish and 28 German learners of English, Biskup (1992) found that German speakers tended to use more creative language and tried to complete the task even if the answer was incorrect. Polish speakers, on the other hand, produced more correct collocations than German speakers but often did not produce anything if they were unsure of the correct translation. It was speculated that the difference in the educational system was the reason for the variance in production between the two groups.

Although test-takers may use strategies to avoid producing a collocation, a productive task is sensitive to the collocational knowledge of the test-taker. After 34 German participants completed a translation task, it was concluded that paraphrasing was not always an easy way to avoid producing a collocation (Bahns et al., 1993). In fact, Bahns et al. found that collocation errors do not differ significantly between good and bad translations. In addition, they observed that translation of verbs was more problematic than translation of other lexical items.

A translation study involving 275 Greek learners of English across eight collocation subtypes in 10 translation items found that grammatical collocations were easier to translate than lexical collocations (Gitsaki, 1999). Lexical collocations in this study included verb-noun collocations as one of the types. A drawback to the interpretation of the findings in this study can be found in the fact that only a limited number of collocations from each type were tested in the translation task, and therefore, the chance of unknown collocations was higher.

Similar to a translation task, the experimental sentence elicitation procedure in Schmitt's (1998) study presented test-takers with a single word and three contextualized prompts in which to use the word. Although test-takers produced 77% of the possible sentences, there were many issues regarding scoring the responses that needed to be resolved before the task could be used effectively. Some sentences did not contain the target collocation. Other test-takers produced more than one collocation in a single sentence. Schmitt recommended that using corpora would be a good way to develop norming criteria to eliminate discrepancies in judgments by native speakers, which are often difficult to reconcile. He also stated that norming lists could be used to produce possible answers and inform the design of a study or test. From another perspective, response analysis might also be used to develop norming lists by analyzing the responses on a constructed response task.

The findings from translation and sentence elicitation studies indicate that although test-takers use strategies to avoid producing a collocation, they do not always produce a collocation. A second finding is that the verb is more difficult for the test-taker to produce than the noun in the verb-noun word pair. This would suggest that the noun is the node in the collocation and the verb would be the collocate. The node is usually the meaning-bearing constituent in a collocation. The collocate is the secondary constituent, which may lose its basic meaning when paired with a node. This is the intuitive approach taken by designers of cloze-type items, wherein the verb is the missing word replaced by a gap in the sentence.

Translation tasks are subject to avoidance strategies. The cloze-type item has the advantage of targeting a specific collocation, which cannot always be done in a translation task yet is still subject to strategies such as L1 transfer, possibly resulting in a negative influence on a test-taker's performance. Nonetheless, gap-filling or cloze-type tasks provide

a different perspective on the data produced due to their limited production format. A constructed response format should produce less variation in test-taker responses, making an analysis of the responses easier. Aghbar (1990) conducted a study comparing 97 ESL students and 44 American students on a 50 verb-noun item test using a sentential cloze format with one gap per sentence and no apparent connection between each sentence in the test. The ESL students produced very few correct collocations. It was speculated in this study that the low number of correct collocations by the ESL students resulted from not knowing the verb-noun pairs rather than lack of knowledge of individual lexical items, given that all of the possible verbs in the collocation pairs were high-frequency verbs. After analyzing the responses from a translation and cloze test, Farghal et al. (1995) suggested that L2 learners tend to select individual items to form a collocation based on free choice and are not aware of a collocation as a multiword unit. Such observations may be more difficult to observe in responses of greater length.

Test-takers are often unaware of why they selected words as combinations. Response analysis can reveal strategies that were used by test-takers. For example, an analysis of responses with translations in a test-takers first language may reveal evidence of first-language (L1) transfer. This is an issue raised in many of the studies as a potential source of error of collocational knowledge. L1 transfer occurs when the target language produced is a direct translation from the first language and does not match the norms of the target language. Evidence of L1 transfer was found in a number of studies (Aghbar, 1990; Farghal et al., 1995; Nesselhauf, 2003, 2005, Voss, 2008). Farghal (1995) observed instances of transfer in about 10% of the incorrect collocations by Arabic learners of English. Nesselhauf (2003) suggested that the influence L1 transfer has on combinations of words may account

for more errors than the previous studies have indicated. Looking at the verb-noun subtype, Bahns (1993) suggested that contrastive analysis is beneficial so that textbook writers and language teachers are informed as to which collocations would be most beneficial for instruction to a particular group of learners with the same L1. This statement comes from an interpretation of the data showing that large a number of collocations have equivalents in English and German. Bahns postulated that German speakers will have little trouble learning and using collocations in English that are equivalent in German, thereby identifying collocations that need not be taught. In this case, studies that restrict the type of collocation and focus on one language family might better be able to support or refute the claim about the usefulness of contrastive analysis in teaching and testing collocations.

It may be worth knowing when L1 transfer occurs in production of language using collocations. These collocations may be less marked for the language learner and require more attention in instruction. To eliminate variation in L1 transfer errors, I recruited participants with a single L1, Chinese-learners of English. However, due to time and space limitations, this paper does not go further to analyze the incorrect responses for L1 transfer or other patterns.

2.5 Scoring

Responses on short constructed response items can be scored as either correct/incorrect (i.e., dichotomously) or according to a rubric awarding scores based on their level of partial correctness (i.e., polytomously). For a number of studies (Bonk, 2001; Durrant, 2008; Eckmans, 2009; Gyllstad, 2009; Revier, 2009), a dichotomous scoring method has been used to distinguish successfully among groups with various proficiency

levels on both receptive and productive collocation tests. This method awarded full credit to a response that matched the collocate that was part of the pre-determined target collocation. The pre-determined collocations were selected based on the sampling method, e.g. intuition, classroom materials, and/or frequency based.

Polytomous scoring methods have been applied to open-response formats (Schmitt 1998) as well as limited-response formats (Aghbar & Tang, 1991). Open-response items require manual analysis of the responses to identify potential collocations and then to determine if the collocation is used appropriately in context. Partial credit may be awarded to combinations that are verified as collocations but not used appropriately in context. Combinations are identified by expert judgment or using a norming list. Expert judges may disagree, however, and pre-determined norming lists will not include all potential verb-noun combinations that may appear in free responses. In addition, more than one collocate may be possible as a response. A polytomous scoring method was used by Schmitt (1998) to measure responses on a 3-point scale, awarding one point to each sentence that included an appropriate collocation. Each prompt required three responses to earn full credit. This means that a test-taker would need to provide multiple word combinations for each prompt, indicating a deep knowledge of each node and their collocates.

Aghbar and Tang (1991) took a different approach to developing a partial credit scale using limited-response items basing correct responses on a four-point scale. Manual analysis determined the degree to which the response was idiomatic, semantic, in the proper register, or combinations of these characteristics. After a comparison of dichotomous and polytomous scoring methods, the authors conclude that a partial credit scale is more rational theoretically and more desirable psychometrically than a dichotomous scale, because a partial credit

scheme has better discrimination among levels. Both partial credit score methods require manual analysis, which is subject to human interpretation regarding idiomaticity and semantic characteristics.

2.6 Summary

In summary, the purpose of early studies of collocational knowledge utilized elicitation tools for research in order to demonstrate a lack of knowledge of collocations by L2 learners of English and to show a large difference in knowledge between L2 learners and target language speakers. Tests of collocational knowledge have not been developed or used as part of a language placement test.

Furthermore, the sampling process for previous studies was unsystematic and sampled collocations from an unknown language domain rather than attempting to reflect a specific domain of language use. The selection process in previous research is not well documented, and the procedures for how the items were selected, in most cases, are not transparent. Moreover, this general sampling procedure makes it difficult to generalize and extrapolate to real-life situations and contexts beyond the test task. Recommendations have been made to teach and test collocations in a particular context or register. Corpus tools have been used to identify collocations in a target language corpus; yet, the frequency of collocations has been based on individual lexical items instead of as a multiword lexical item. The selection of collocations for a test of collocational ability should be based on the identification of collocations as single units identified by frequency in a particular target language domain.

In addition to a variety of sampling methods, tasks have varied as well. Elicitation formats have been mostly experimental. Selected response formats are not often contextualized and may be measuring something other than collocational knowledge. Item types that allow a free response are subject to strategies which assist the test-taker in avoiding the use of target collocations. A contextualized gap-filling item type with a sentential context may narrow the use of strategies to those which might help the test-taker produce a target response. Response analysis may also uncover certain strategies that are used by the test-takers and better discriminate among groups using polytomous scoring.

2.6 Conceptual framework

The issues of test purpose, target language domain, item selection, and task format have been central to the decisions that were made in the development of the collocational ability test for this study and, therefore, are included in the test framework. This theoretical framework helped define concepts that are used to design, develop, and evaluate the tasks for the test of collocational ability.

2.6.1 Construct definition

The construct to be assessed is "collocational ability in academic written texts." The frame of reference for this construct is based on the interactionalist definition (Chapelle, 1998), whereby "*performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts*" (p. 43 italics in original). The trait in this case is knowledge of collocations in the target language context, which is written, academic English. Performance on a language test is thus

an indicator of collocational knowledge which is produced in context whereby it can be assumed that the performance is a sample of such a performance in other contexts. Furthermore, "an interactionalist construct definition comprises more than just trait plus context; it includes the metacognitive strategies (i.e., strategic competence) responsible for putting person characteristics to use in context" (p. 44). The language user draws on metacognitive strategies to assess the context and produce appropriate language (e.g., acceptable collocation) that is appropriate for the context (e.g., written academic English). The construct of collocational ability based on the interactionalist perspective would include the knowledge of collocations and the processes needed to produce contiguous or non-contiguous collocations and the context in which they are produced as well as the metacognitive strategies to direct and assess their use.

Interpretations of score meaning from the collocation test include knowledge of collocations as one part of the interactionalist construct definition. For this test, the term *collocation* is the same as the definition of a collocation identified by linguists as the occurrence of two or more words within a short space of one another in a text (Sinclair, 1991). Similarly, Nation (2001) defined *collocation* more specifically, stating that

the term 'collocation' is used to refer to a group of words that belong together, either because they commonly occur together like *take a chance*, or because the meaning of the group is not obvious from the meaning of the parts, as with *by the way* or *take someone in* (trick them) (p. 317).

For the purpose of this test, the collocations are a combination of both contiguous and non-contiguous verb-noun restricted collocations. Collocations are restricted in the number of collocates that can form combinations with the node. Some nodes have a large number of

collocates, whereas others may have relatively few. Words that can be combined freely are not restricted.

Restricted collocations are word combinations that are restricted in their commutability where the meaning of the word combination is made up of the sum of its constituents (Aisenstadt, 1979). This description follows the phrase-based definition of collocations, which considers a collocation as a multiword lexical item. Aisenstadt (1981) outlined the three criteria which are used to identify these multiword lexical items, also known as restricted collocations: (1) their structural patterns, (2) the commutability restrictions, and (3) the meanings of components. A detailed taxonomy of collocations is also provided by Howarth (1998). Figure 2.1 shows the parts of Howarth's collocation continuum.

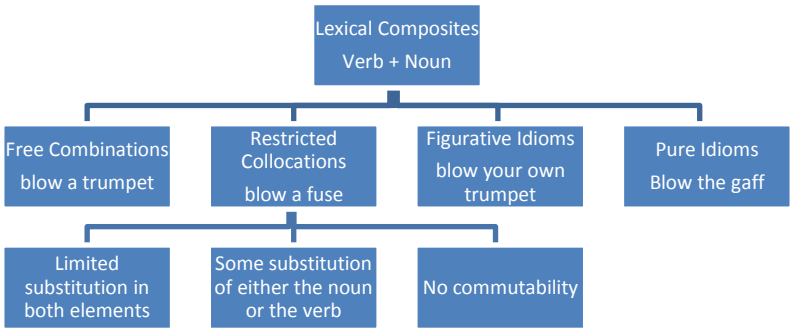

Figure 2.1. Howarth's (1998) collocation continuum (p. 28)

Howarth's (1998) collocation continuum provides detail and depth to Sinclair's (1991) distinction between the open choice principle as free combinations and the idiom principle as pure idioms, all of which are combinations of words or lexical composites selected to create meaning. The second row in Figure 2.1 describes the combinations between the two principles based on semantic analysis where the interpretation of the

meaning of words combined freely is more transparent than the interpretation of the meaning of words combined according to the idiom principle. The meaning of each word in the free combination "blow a trumpet," for example, is understood individually, whereas the first constituent "blow" in "blow a fuse" has a different meaning than the core meaning of "to blow." The intended meaning of this multiword lexical item is different than the combination of the intended meaning of each individual lexical item. These are considered restricted collocations because the number of acceptable constituents that may be substituted in the combination is limited, termed *commutability*, as presented in the third row. The number of substitutions of either the noun or the verb is limited and thus restricted in word combinations. The term *restricted collocation* is related to the phrase-based approach to collocation in studies by Revier (2009), which consider a collocation as a single lexical item. The construct includes the process of producing one constituent in a multiword lexical unit as evidence of knowledge of the word combinations as a single unit or phrase.

Restricted collocations in this study are of the verb-noun type with restrictions on the commutability of the verb. Moreover, one of their constituents has a different meaning than the meaning of the single lexical item, causing a shift in meaning. These two criteria place the collocations somewhere in the middle of the continuum, between free combinations and the item principle. In this study, the term *collocation* refers to a verb-noun restricted collocation type which is regarded as a single contiguous or non-contiguous lexical unit.

2.6.2 Context of academic discourse

The interactionalist construct definition also includes the target context or a target language domain in which the language knowledge is used. The context for collocation use

about which the test scores are relevant is written academic English at an English-medium institution of higher education. The collocations have, therefore, been identified in large target language corpus using a corpus-driven method to sampling.

Since the purpose of the collocation test is to develop an instrument that can elicit a score that can be used as an indicator of academic English collocational ability and score use which can contribute to admission and placement decisions at English-medium institutions, we look for distinctive characteristics in a large sample of general academic language. In a language domain, collocations are fairly infrequent and require a large amount of language in order to identify frequently occurring verb-noun restricted collocations that will be representative of general written academic language. To accomplish this, a large collection of texts was used in this study to identify frequent collocations that appear in written academic language. The texts that were used are the academic files of the British National Corpus (BNC), identified by Lee (2001), which provide a large number of texts from which the target language use (TLU) domain can be investigated as a sub-corpus. This written academic sub-corpus of the BNC consists of approximately 16 million running words and was the largest collection of written academic texts at the time of this test development.

The language in the academic written sub-corpus of the BNC represents texts that students will encounter in a course at an institution of higher education. It is also the language that students should be able to produce by the end of their coursework in order to participate as professionals in their field. Although there is variation in the use of certain collocations based on variety of English, high-frequency collocations in an academic domain have been selected from a corpus of academic English that is representative of professional

academic language as found in institutions of higher education and not bound by a particular variety of English.

The scoring scale is related to the raw frequency of both constituents in the collocation pair as identified in the academic files of the BNC corpus. Similar to the phraseological approach in the study by Revier (2009), I have identified items for this study based on the frequency of both constituents in the verb-noun collocation as a single lexical unit, without predetermined single lexical items. This phrase-based approach to identifying collocations as lexical items has been suggested by Shillaw (2009) to provide stronger evidence of collocational ability based on the selection process. This is a different perspective from the traditional method of corpus identification using a word word-list approach, whereby base words were then selected from the high-frequency list and used as nodes to identify collocates using corpus tools (see Gyllstad, 2005; Mochizuki, 2002; Moreno Jaén, 2007).

When a list of words is selected before the second word in the collocation is identified, there is early intervention by a human on the selection process which influences the items on the test. Delayed or late human intervention is a manual analysis of an automatic search after a list of word pairs has been identified by the software. Theoretically, tests with items selected using early human intervention would measure the depth of vocabulary knowledge for the initial list of base words rather than the knowledge or ability of the collocations as a lexical unit. This word-based approach has limitations and raises questions as to the validity for the definition of a *word* as a single or multiword lexical item (see Gardner, 2007 and Shillaw, 2009).

Although Revier's (2009) approach identified word pairs based on frequency of both constituents, there were two drawbacks to this approach. First, only contiguous word pairs were identified. This procedure limits the selection to a specific syntactic form, thereby limiting the possibility of the frequency of a word pair with a different word form that has a different frequency position. Secondly, items were selected only if they contained one of 15 pre-selected verbs, which is a form of early human intervention. The pre-selection of a list of verbs requires early human intervention and is similar to the word-based approach, whereby base words are selected before the entire collocation, which would limit the validity of the "whole collocation" approach.

The sampling method in this study is based on high-frequency target collocations that are identified in a large target language corpus using corpus tools. The collocations are identified as multiword lexical items with late human intervention.

2.6.3 Metacognitive strategies

The third part of the interactionalist construct definition consists of metacognitive strategies that are used to direct and assess the use of collocations in context. This is also part of what Bachman and Palmer (2010) refer to as strategic competence.

A number of metacognitive strategies have been identified and studied in previous research on collocational knowledge. Howarth (1998) has categorized strategies used by L2 learners as (1) avoidance, which includes paraphrasing and synonymy; (2) experimentation, or taking risks to combine words freely; (3) first-language transfer; (4) analogy, which is overlapping or blending known collocations; and (5) repetition, using high-frequency verbs.

Some of these strategies may assist a test-taker in producing an appropriate collocation, whereas others prevent the test-taker from producing a target response.

In a free production task, a test-taker can use a variety of avoidance strategies such as circumlocution, paraphrasing, changing the topic, or semantic avoidance (Blum-Kulka & Levinson, 1983). Evidence has been found for the use of strategies such as avoidance or paraphrasing as well as producing language that contains free combinations rather than produce incorrect word combinations (Bahns & Eldaw, 1993; Farghal et al., 1995; Howarth, 1998). Synonymy and L1 transfer are also common strategies used by test-takers, which result in collocations that are either not acceptable in English or not appropriate for the context (Farghal et al., 1995; Voss, 2008).

Transfer and repetition are two metacognitive strategies that have potential to assist a test-taker in producing a target response. If the collocation is similar in a test-taker's first language, L1 transfer may be beneficial. In addition, many collocations contain high-frequency verbs, sometimes called *light verbs*, such as *have*, *make*, and *take*. Metacognitive strategies may be helpful to the L2 learner in communicating in real life but make measurement of collocational ability more difficult. Strategies that assist an L2 learner to communicate in real-life situations are not always desirable in a test setting. A successful paraphrasing strategy, for example, may result in a response that does not contain a target collocation, and as a result, knowledge of a particular target collocation by that L2 learner remains unknown. Measurement instruments can only measure a test-taker's ability if there is evidence. Task format on a test is critical in prompting particular metacognitive strategies.

2.6.4 Relationships with other constructs in nomological network

A construct is defined for a particular test in accordance with a language ability of interest; yet, even though a construct is "constructed" to include language abilities related to the purpose of the test, the language ability of interest still has theoretical connections with other constructs. We cannot define the construct for a reading test without acknowledging that the amount and degree of vocabulary knowledge by a test-taker is related to the test-taker's reading ability. We must also recognize that knowledge of word combinations by a test-taker also relies, to some degree, on the knowledge of individual words, assuming that the word pairs have not been learned as whole units. This "interlocking system" of theoretical relationships among various constructs is referred to as a *nomological network* by Cronbach and Meehl (1955). One way to observe a relationship among underlying language ability theory is to compare test components using correlation coefficients (Alderson, Clapham, & Wall, 1995). Nomological evidence is made possible by observing patterns of relationships between test scores and predicted theory. The use of a theoretical foundation would also be known as a strong program, as suggested by Cronbach and Meehl (1955) and Meehl and Golden (1982), which restricts the types of correlations that are allowed as evidence in the validity argument (Kane, 2001). By contrast, the weak program does not limit the types of comparison and, thus, is open to rebuttals reducing the claims in the validity argument. Theoretical relationships among other constructs related to collocational ability can be predicted. The correlational coefficient could then be used as evidence to support the extrapolation inference in the interpretive argument and used as a valid indicator of academic language performance in an academic setting.

Performance on vocabulary tests has been clearly established as generally having a strong relationship with performance on reading comprehension tests (Alderson, 2000). The relationship between reading ability and vocabulary has few skeptics; however, the traditional view of vocabulary has been on the recognition of individual words. The correlation between reading ability and vocabulary size has been quantified and linked to the level of relative frequency of single words in use today. For instance, there is a positive relationship between the number of frequent words that are known and higher performance on reading ability. Research on the relationship between vocabulary size and reading has found that lexical level of an L2 learner is an even better predictor than a measure of general proficiency (Laufer, 1992). Comparing scores on a test of general vocabulary and a test of reading comprehension, Johnston (1984) reported a correlation coefficient of 0.35. Grabe (2009) cited several studies with correlations between vocabulary and reading at $r = .63$ and above in L2 settings. The amount of correlation is also related to the degree of specialized vocabulary. General vocabulary measures tend to correlate less than vocabulary in a particular context.

Vocabulary difficulty can also have an effect on the reading comprehension for both first- and second-language readers (Alderson, 2000). A study of problems associated with reading in an L2 at the university level in Norway claims unfamiliar vocabulary as a major hindrance (Hellekjaer, 2009). Problems with understanding a passage may arise, however, even if all of the individual words are familiar to a reader. Certain word combinations have been shown to impede reading comprehension despite the fact that all individual words are likely to be known (Martinez, 2010). Thus, other aspects of vocabulary knowledge such as collocation may provide information about performance on reading comprehension tests.

Vocabulary depth and breadth are interconnected and interdependent components of vocabulary knowledge. Qian (1999) confirmed the predictive value of scores on a vocabulary size test and general academic reading comprehension but also suggested that scores on a depth of knowledge test, including word associations and collocation, also can be a beneficial predictor. These are promising claims even though they are based on measures of receptive vocabulary knowledge. The collocational ability test is a productive ability task. The type of depth of knowledge that corresponds to the test is verb-noun collocation that has a restricted quality that carries certain semantic properties that are lost when the words are co-selected. The noun or node usually retains its core meaning while the verb or collocate becomes less lexicalized when co-selected with the node. "Catch a ball," for example, retains the sense of each word, but in the more restricted collocation, "catch a cold," the meaning of *catch* loses its most common meaning: "to intercept and hold (something that has been thrown, propelled, or dropped)" ("Catch," 2010, def. v.1). Despite the loss of semantic meaning of the collocate, however, such restricted collocations are not difficult for L2 learners, because the node retains its core meaning. The collocates are often overlooked and not retained once they are encountered; therefore, reading comprehension is not lost entirely by the presence of such restricted collocations. Due to the productive nature of the task on the collocational ability test, one might conclude that a positive relationship will exist between performance on a reading measure and performance on the test of collocational ability because of the relationship between the knowledge and use of single words that are necessary in both contexts and familiarity with their combinatory principles. Due to the nature of the verb-noun restricted collocations, however, reading comprehension is not impeded a great deal by the presence of such collocations. Theoretically, scores on the collocational ability test may have

a weaker relationship with scores on a test of reading ability than scores on a vocabulary size test.

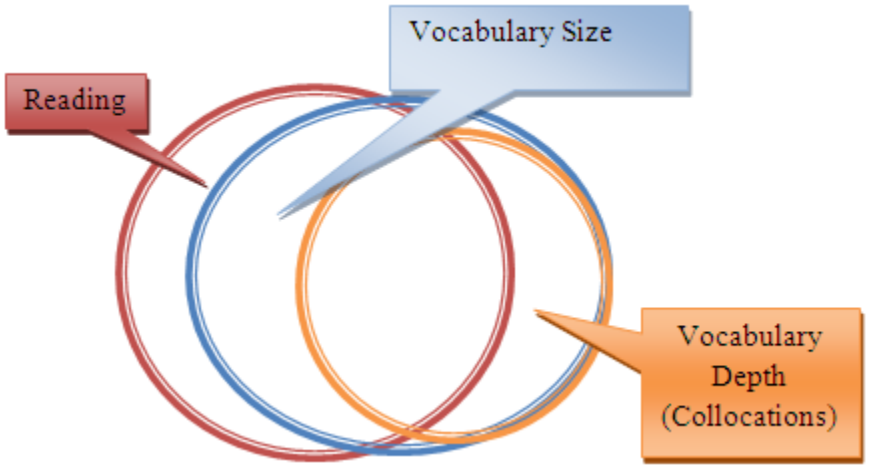

Figure 2.2. Theoretical relationships in nomological net

The relationship among the constructs in a nomological net might be represented in a formula such as the following.

R/VS > VS/VD > R/VD

One would predict the relationship between reading (R) and vocabulary size (VS) to be stronger than the relationship between vocabulary size (VS) and vocabulary depth (VD). Consequently, one would also predict the relationship between reading and vocabulary depth, represented by collocational ability in this study, to have the weakest relationship.

2.7 Response formats

In view of the intention to measure an interactionalist construct of collocational ability, the Collocational Ability Test was developed using limited constructed response items. This item type was selected because of its successful use eliciting target collocations in other studies. A variety of tasks have been trialed to elicit either one or both parts of a verb-noun combination, including multiple choice (Keshavarz & Salimi, 2007; Koya, 2003, Moreno Jaén, 2007), gap-filling (Bonk, 2001; Aghbar, 1990; Aghbar & Tang, 1991), matching (Gyllstad, 2006), sentence elicitation (Schmitt, 1998), and translation (Koya, 2003). As mentioned above, some strategies such as paraphrasing can be beneficial for language production, yet their use makes the elicitation of a target collocation difficult in a free production task. To eliminate the use of strategies that inhibit a test-taker from producing a target response, the task must be designed with a limited response format.

Translation and gap-filling are productive tasks that are susceptible to the use of meta-cognitive strategies such as the avoidance techniques mentioned above. On the other hand, the gap-filling item type, or cloze item type, is a limited production task which can be designed to elicit a more specific target response such as the collocate in the target collocation. A sentence with a short one-word response in a gap to provide a verb as the collocate in a verb-noun collocation would eliminate the use of avoidance or paraphrasing strategies but will not, however, eliminate other strategy use including, experimentation, transfer, analogy, or repetition. Response analysis could uncover the use of transfer, analogy, or repetition. Furthermore, observations about strategy use can be confirmed or rejected with qualitative data collection though interviewing test-takers and capturing performance during test administration. This type of data collection was attempted in this study to confirm the

results from the quantitative data analysis. In light of the perceived relationship with metacognitive strategy use, a gap-filling (cloze) item type was chosen for this test to elicit a single lexical item as part of a multiword item and to prevent the use of avoidance strategies while providing the context needed to support the interactionalist approach in language testing.

2.8 The interpretive argument

Score interpretation is based on an interpretive argument, which is the first step in developing a validity argument (Bachman, 2004; Kane 1992, 2001; Mislevy, Steinberg, & Almond, 2003). The interpretive argument defines the grounds, inferences, warrants, and claims that provide the foundation for score interpretation. The interpretive argument can be used before the test is operational, during the test design process, as well as in the development of the validity argument. The inferences that underlie score interpretation are domain description, evaluation, generalization, explanation, extrapolation, and utilization. Domain description links performances of collocational ability in the target domain with performance in the test domain. The evaluation inference provides observed scores that reflect collocational ability. The generalization inference claims that expected scores represent consistent performance across items. The explanation inference attributes the expected scores to a construct of collocational ability in academic writing. The extrapolation inference links expected scores with test performance beyond the test to the academic domain. The utilization inference links score with decisions that are made based on the scores.

Each inference requires evidentiary support for each of the assumptions on which the inference is based. Warrants and assumptions are identified for each inference, describing the type of research necessary to provide backing for each assumption. After the test is operational, additional research can provide support for the assumptions in each inference.

The conclusion for the interpretive argument underlying score use is that the score is useful as one piece of information contributing to placement decisions of L2 learners of English in English-medium colleges and universities.

Figure 2.3 below outlines the interpretive argument that underlies score interpretation as an indicator of academic English collocational ability and score use to be part of admission and placement decisions at English-medium institutions. In other words, the claim that the interpretive argument needs to support is that the test score reflects the ability of the test-taker to understand and use restricted verb-noun collocations as they are written in college and university settings.

Based on the TOEFL interpretive argument (Chapelle, Enright, & Jamieson, 2008), the structure of the seven-part argument in the interpretive argument for the Collocational Ability Test is shown in Figure 2.3. The grounds for this interpretive argument begin with the domain of academic English use. Because the test is still in the development stage, backing for the assumptions underlying the utilization and impact intention inferences is not possible.

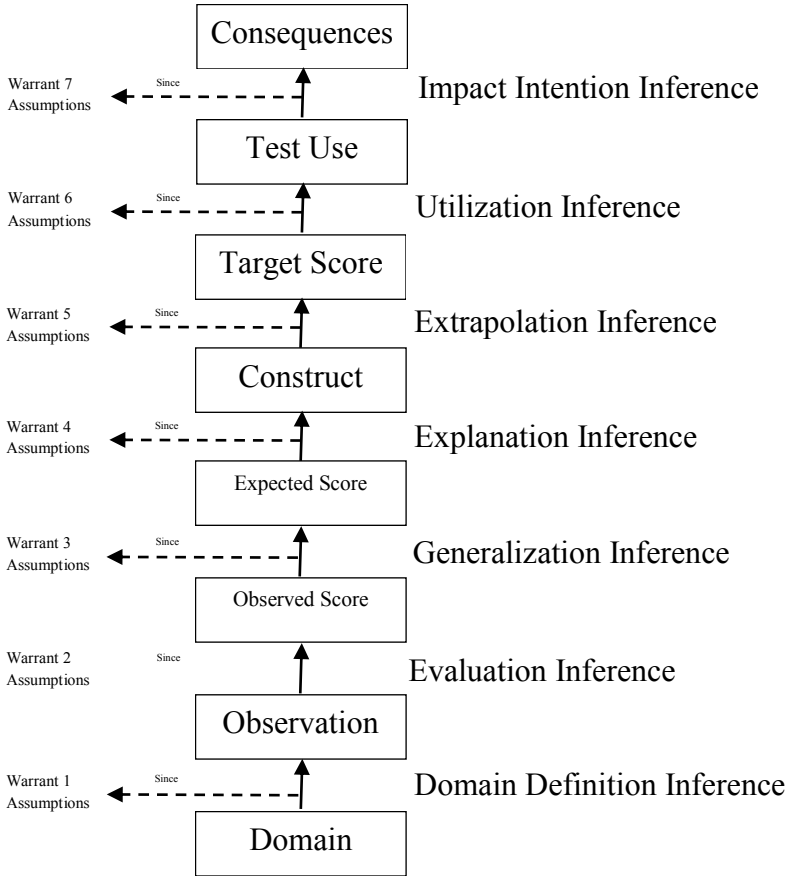

Figure 2.3. Interpretive argument structure for the Collocational Ability Test

The first part, the domain definition inference, is based on the warrant (1) in Figure 2.3) that observations of performance on the collocation test reflect the collocational ability that is relevant and representative of the target domain of language use in English-medium colleges and universities. This warrant assumes that:

1. Performance in written academic English tasks relies in part on ability to use and understand verb-noun collocations.
2. Collocations that appear on the test were selected from a large corpus of written academic discourse.
3. Appropriate collocations for the collocation test have been identified.

4. Test tasks elicit performance from test-takers that reflect their collocational ability.

The evaluation inference is based on the warrant (2) that observations of performance on the collocation test are evaluated to provide observed scores that reflect collocational ability. This warrant is based on six assumptions. Research question (RQ) 1 was developed based on the sixth assumption.

1. The scoring procedure captures the collocational ability learners have related to word pairs based on frequency of the whole collocation.
2. The scoring procedure accurately captures the varying degrees of strength of ability for collocations that learners have as they develop collocational ability.
3. The acceptance of misspelled words results in a score that reflects collocational ability (and not spelling ability).
4. Task conditions allow test-takers to perform in a manner that exhibits what they know.
5. The statistical characteristics of items and measures are appropriate for norm-referenced decisions.
6. Items on the test fit the model predicted by a Rasch IRT analysis and measures are appropriate for distinguishing among test-takers (RQ 1).

The generalization inference is based on the warrant (3) that observed scores are estimates of expected scores, which are comparable across the items on the test. Research question 2 is based on the first assumption in the generalization inference. This warrant is based on the assumptions that:

1. Estimates of test-takers' performance can reliably distinguish among test-takers (RQ 2).
2. The tasks and test specifications are sufficiently detailed and consistent to produce equivalent test forms.
3. The computer-based administration of the test is sufficiently uniform to produce consistent results.

The explanation inference is based on the warrant (4) that expected scores are attributed to a construct of collocational ability in academic writing, which includes knowledge of whole collocations. Three research questions are based on assumptions 3, 4, and 5. The five assumptions are:

1. Performance on the collocation test reflects test-takers' collocational ability.
2. The construct of a restricted lexical collocation has been defined as a whole collocation rather than individual constituents.
3. The more frequent a collocation, the easier the corresponding item will be for test-takers (RQ 3).
4. The scores on the Collocation Ability Test correlate as predicted to other tests of English ability related to the construct (i.e., reading and a productive vocabulary size) (RQ 4).
5. While taking the test, test-takers use metacognitive and cognitive strategies related to collocation use in academic language (RQ 5).

The extrapolation inference is based on the warrant (5) that the construct of collocational ability as assessed by the Collocation Ability Test accounts for the academic

collocational language performance in academic discourse in English-medium colleges and universities. Two research questions (RQ 6 & 7) come from the assumptions that:

1. The collocations appearing on the test reflect those that the test-takers will find in an academic context.
2. Scores on the collocation test distinguish among proficiency groups with and without experience and knowledge of academic language (RQ 6).
3. Scores on the collocation test have a positive relationship with scores on other measures of academic vocabulary (RQ 7).

The utilization inference is based on the warrant (6) that performance on the test contributes to making appropriate decisions about matriculation and placement in English-medium colleges and universities. The assumptions that underlie this warrant are that:

1. Score-based interpretations provide enough information to contribute to the decision-making process.
2. Test scores contribute to and facilitate student placement in English language courses at English-medium colleges and universities.

Finally, the impact intention inference is based on the warrant (7) that test score interpretation and use is beneficial for all test users and stakeholders. The two assumptions are:

1. The test construct raises awareness about the importance of collocations in academic English.
2. Instructors are aware of the potential benefits of alternate scoring methods for constructed response tasks.

In order to collect evidence for the last two inferences and provide backing for the assumptions to support their underlying warrants, the test should be operational. Whereas this study is a test development project, and the test is not yet operational, backing cannot be collected at this time.

The development of the research questions was guided by the potential strength of the assumptions underlying the warrants that support the inferences in the interpretive argument. The assumptions that are perceived to be weakest and need the most backing were candidates to be included in research questions. The research questions that were developed are presented in the next section. A summary of inferences, warrants, and assumptions for the interpretive argument is presented in Table 2.1. The assumptions on which the research questions are based are marked with the research question number in parentheses after the assumption.

Table 2.1. Summary of inferences, warrants, and assumptions for the interpretive argument

Inference	Warrant licensing the inference	Assumptions underlying inferences
Domain description	Observations of performance on the collocation test reflect the collocational ability that is relevant and representative of the target domain of language use in English-medium colleges and universities.	1. Performance in written academic English tasks relies in part on ability to use and understand verb-noun collocations. 2. Collocations that appear on the test were selected from a large corpus of written academic discourse. 3. Appropriate collocations for the collocation test have been identified. 4. Test tasks elicit performance from test-takers that reflect their collocational ability.
Evaluation	Observations of performance on the collocation test are evaluated to provide observed scores that reflect collocational ability.	1. The scoring procedure captures the collocational ability learners have related to word pairs based on frequency of the whole collocation. 2. The scoring procedure accurately captures the varying degrees of strength of ability for collocations that learners have as they develop collocational ability. 3. The acceptance of misspelled words and variation in word forms results in a score that reflects collocational ability. 4. Task conditions allow test-takers to perform in a manner that exhibits what they know. 5. The statistical characteristics of items and measures are appropriate for norm-referenced decisions. 6. Items on the test fit the model predicted by a Rasch IRT analysis and measures are appropriate for distinguishing among test-takers (RQ 1).
Generalization	Observed scores are estimates of expected scores, which are comparable across the items on the test.	1. Estimates of test-takers' performance can reliably distinguish among test-takers (RQ 2). 2. The tasks and test specifications are sufficiently detailed and consistent to produce equivalent test forms. 3. The computer-based administration of the test is sufficiently uniform to produce consistent results.
Explanation	Expected scores are attributed to a construct of collocational ability in academic writing, which includes knowledge of whole collocations.	1. Performance on the collocation test reflects test-takers' collocational ability. 2. The construct of a restricted lexical collocation has been defined as a whole collocation rather than individual constituents. 3. The more frequent a collocation the easier the corresponding item will be for test-takers (RQ 3). 4. The scores on the Collocation Ability Test correlate, as predicted, to other tests of English ability related to the construct (i.e., reading and a productive vocabulary size) (RQ 4). 5. While taking the test, test-takers use metacognitive and cognitive strategies related to collocation use in academic language (RQ 5).
Extrapolation	The construct of collocational ability as assessed by the Collocation Ability Test accounts for the academic collocational language performance in academic discourse in English-medium colleges and universities.	1. The collocations appearing on the test reflect those that the test-takers will find in an academic context. 2. Scores on the collocation test distinguish among proficiency groups with and without experience and knowledge of academic language (RQ 6). 3. Scores on the collocation test have a positive relationship with scores on other measures of academic vocabulary (RQ 7).
Utilization	Performance on the test contributes to making appropriate decisions about matriculation and placement in English-medium colleges and universities.	1. Score-based interpretations provide enough information to contribute to the decision making process. 2. Test scores contribute to and facilitate student placement in English language courses at English-medium colleges and universities.
Impact intention (Consequences)	Test score interpretation and use is beneficial for all test users and stakeholders.	1. The test construct raises awareness about the importance of collocations in academic English. 2. Instructors are aware of the potential benefits of alternate scoring methods for constructed response tasks.

2.9 Research questions

The research questions were developed to direct attention to inferences in the interpretive argument that may require particular attention. Backing for the domain definition and evaluation and part of the explanation inferences can be found in the research and scoring procedure explained in the framework. The research questions are addressed using data from both the dichotomous and polytomous scales. The following research questions were developed to guide the empirical research used to provide backing needed for part of the generalization, part of the explanation, and the extrapolation inference. The research questions were developed based on assumptions that needed backing. The corresponding inference and assumption for each research question is listed in parentheses.

1. Do items on the test fit the model predicted by a Rasch IRT analysis? (evaluation: assumption 6)
2. Does the test constructed using high-frequency collocations in general academic English distinguish among groups with acceptable reliability? (generalization: assumption 1)
3. Do the means of the item facilities correspond to what would be predicted by their rank order in the frequency list? (explanation: assumption 3)
4. Do test-takers' performance on the collocation test correlate positively to performance on a concurrent measure of English ability as expected? (explanation: assumption 4)
5. Does student's perception of and experience with academic collocations and academic language match their performance on the collocation test? (explanation: assumption 5)

6. Do test-takers at different proficiency levels perform differently on the collocation test based on their experience and knowledge of academic language? (extrapolation: assumption 2)
7. Is there a strong correlation between scores on the collocation test and scores on the academic vocabulary sub-test as predicted? (extrapolation: assumption 3)

2.10 Chapter summary

This chapter reviewed the literature that is relevant to and helped guide the work for this dissertation. It provided a review of previous and current research measuring collocational knowledge. Issues relevant to the study included an overview of test purpose, identification of the target language domain, sampling collocations, development of the task, and scoring methods. The chapter introduced the conceptual framework that assisted in making informed decisions regarding the design and evaluation of the Collocational Ability Test. Finally, the interpretive argument and the research questions were introduced. Chapter 3 describes the issues related to development and trialing of the Collocational Ability Test.

Chapter 3 Test development

This chapter describes the development of the Collocational Ability Test. The test was developed following the framework described in chapter 2. A summary of task characteristics and variables can be found in Appendix A. The following sections describe steps in the test development process as they relate to the inferences in the interpretive argument for the Collocational Ability Test.

3.1 Development relevant to domain description

The domain description is based on the inference that observations of performance on the collocation test reflect the collocational ability similar to those in the target domain of language use in English-medium colleges and universities. The meaning of a score on the test is intended to represent collocational ability in a target domain; thus, selecting a corpus to represent the target language domain is important for sampling appropriate language features. This section describes the selection of the corpus and explains the sampling procedure. It concludes with a discussion about the design of the gap-filling task.

3.1.1 TLU Corpus Selection and Description

The corpus that represented the domain of academic written English for this study was a sub-corpus of the BNC. The corpus was selected rather than developed specifically for this study. The written academic files of the BNC were relevant and representative of the language needed to identify target collocations for the test. Considerations in selecting an appropriate corpus included the design and construction, balance and representativeness, annotation including text format and encoding, and alternative corpora.

The design and construction is the first consideration when selecting a corpus. Rather than search for texts that include linguistic features, a corpus should be developed for its communicative function rather than the language it contains (Sinclair, 1991). This idea was applied to the BNC when it was developed from a wide range of sources as a cross-section of British English at a particular time period. It is a monolingual, synchronic (late 20th century), general corpus.

Selecting a section of a larger corpus reduces the sample size and has an effect on the analysis of the language sampled; however, the size of the corpus selected depends on the purpose of its use (Hunston, 2002). The size of the corpus is also related to the primary interrogation method, quantitative or qualitative. Because the occurrences of restricted verb-noun collocation pairs are few, a quantitative approach to identifying such pairs requires a large corpus. The instances identified provide evidence for the development of a theory-based test of collocation ability. Cheng, Greaves, Sinclair, and Warren (2009) stated two reasons for using a large corpus. First, it provides a critical mass of instances, and second, theoretical statements can be made due to the large sample size. Although a smaller corpus could be used to develop test items from a specific academic discipline, the number of instances may not be enough to support a generalization to other academic situations. The use of a small specialized corpus is advantageous for qualitative methods and ESP materials development.

The academic files of the BNC were one of a few choices in corpora of general academic English of any substantial size at the time of test development. The general academic files were identified using an index to the BNC developed by Lee (2001). Using an excel spreadsheet the files were identified as academic written English and included in the

sub-corpus. Although the entire corpus is made up of 100 million words, 505 written academic files contain 16,077,495 running words, according to a word count using ConcGram 1.0. This word count is 171,511 more words than calculated using Lee's (2001) BNC Index. The difference in word counts could be attributed to how the software measures compound words and/or punctuation.

The texts were collected between 1991 and 1994 but contain texts that are older; some texts date back to 1975. Permission was obtained and guidelines were followed for sampling from longer texts to prevent the over-generalization of idiosyncrasies from a particular style or topic (Sinclair, 2005).

The second consideration in selecting a corpus is the balancing and representativeness of the texts in the corpus. In addition to the selection criteria, classification features were identified; an appropriate level of variation was sought over fixed proportions of these features. Although these considerations are important to the construction of the corpus, they are not as relevant to the interpretation of the data collected for my research questions. Although the BNC has been developed with samples of English in two modes, written and spoken, my research questions were best answered by looking at the files of written English. The selection criteria for the written texts included domain, medium, and time. The BNC limits the medium (mode) to written or spoken. Other corpora have a finer-grained description including classifications such as "written-to-be-read" and "written-to-be-spoken" texts (Chen, Huang, Chang & Hsu, 1996).

Texts in the corpus are both published and unpublished, consisting of six academic disciplines or genres, according to Lee (2001). These are technical engineering, social science, politics and law, natural science, medicine, and humanities and arts. Aston (2001)

criticized the balance of texts in the BNC, claiming that some types of the written texts are read more often than others and should not be weighted equally. This is a design issue that could influence the description of language use. The representativeness of the texts included in the corpus is an important consideration, because it has an impact on the verb-noun instances identified.

All of the texts in the sub-corpus were written by both male and female adults and collected between 1991 and 1994. The texts include those written by single and multiple authors as well as publications by corporations. The ages of the authors fall into five age groups with a fairly even mix of circulation status. While books and periodicals fit into Biber's (1993) sample frame for published writing, there are a few unpublished manuscripts to represent the unpublished sample frame, due to many of the written texts included in the corpus coming from a publisher.

A third consideration when selecting a corpus is the necessity to annotate, format, or encode the corpus in any certain way. Sinclair's (1991) "clean-text policy" recommends leaving the text in a standard format that is unprocessed and free of other codes or erroneous tagging. Original Constituent Likelihood Automatic Word-tagging System (CLAWS) annotations and tagging were removed from the corpus for use with the software. Sinclair (2005) warned against this process and recommended that text be kept in a uniform format to avoid the time-consuming, laborious task of converting text. The main reason to use plain text, however, is that ConcGram 1.0 interrogates text in plain text format. The tags were removed from the BNC texts using WordSmith Tools 5.0 and converted to Unicode using a batch file and a DOS command.

Finally, to consideration of alternative corpora is always necessary. The BNC has been used to identify verb-noun collocation pairs by other researchers for the development of collocation tests, yet they have always used the entire corpus. A sub-corpus of the BNC was therefore, a logical choice for a quantitative identification of verb-noun collocation pairs in written general academic language. The selection of an academic sub-corpus draws upon a corpus that has been used in similar research but in a more systematic way that supports the sub-corpus as a suitable choice.

Although the selection of the academic texts in the BNC is a suitable choice, one might object to its use for several reasons. The high percentage of published texts in the corpus could be viewed as a disadvantage of the sub-corpus chosen for this study. A more realistic target domain for students in tertiary education might be samples of texts written by proficient student writers of English. Three other corpora were candidates under consideration for this project at the beginning of the test development.

The first alternate corpus, also academic British English, is The British Academic Written English (BAWE), collected from 2004-2007 and containing almost 3000 texts from British Higher Education ranging from 500-5000 words in 35 disciplines. A second corpus, the Cambridge Corpus of Academic English**,** is much larger than the BNC files with 30 million words, which is a collection of samples of both American and British English. Finally, a third potential corpus, The Michigan Corpus of Upper-level Student Papers (MICUSP), is a collection of approximately 2.6 million words from four academic divisions. Although one could argue that one of these three alternate corpora might be considered more appropriate as target language for the collocation test being developed, financial and proprietary issues prevented their immediate use.

One issue relevant to the selection of an appropriate corpus was the differences between British and American English. Other studies do not mention the use of the BNC or differences in varieties of English. Attention to this distinction is necessary to determine frequency counts as well as to identify target collocations for specific language tests uses. Verb-noun pairs that share common usage in both language varieties were found by the software, such as the pair "make ~ decision." Also included in the search results were word pairs such as "take ~ decision," which is less frequently used in American English. A search in the Corpus of Contemporary American English (COCA) revealed the trend of these two word pairs in their basic form. "Make a decision" was found 1421 times where as "Take a decision" was found only 17 times; however, "Make a decision" was the second most frequent collocation, whereas "Take a decision" was 26th. Moreover, the collocation "Take a decision" was eliminated in the test revision. Although variation exists in the use of certain collocations based on variety of English, high-frequency collocations in a particular target language domain have been selected from a corpus of target language text that is representative of professional academic language as found in institutions of higher education.

3.1.2 Identification and selection of target collocations

The second step in the domain analysis was to identify and select collocations relevant to the target domain, select an appropriate task, and develop the test items. This section describes this process and the corpus software that was used. The collocations for the collocational ability test were identified in and selected from written academic English as the target language domain for context rather than an unknown English domain that most studies have used.

The following method was used to identify collocations from a corpus of general academic written discourse, a sub-corpus of the British National Corpus. A corpus-driven approach was taken to identify collocations in the corpus, beginning with an automatic identification of collocation pairs using ConcGram 1.0 software. Several search engines are capable of identifying verb-noun collocation pairs. Some are web-based while others range from free concordancers (e.g., Antconc) to commercial products (e.g., Monoconc, Wordsmith Tools). ConcGram 1.0 was selected because of its ability to identify all 2-word concgrams (without lemmatizing) in a corpus without the use of statistics.

The software was run using Windows XP on a Mac running parallels. A stop list was used to eliminate pairs which include prepositions, articles, and the copula "be." This software produced a list of two word concgrams based on raw frequency within a span of ±4. This list was sorted by raw frequency within a range of 20 as the fewest instances up to 881. A manual analysis was conducted to select collocations that were representative of the collocation type *restricted verb-noun collocation* from highest frequency count to lowest. The most frequent word pairs that fit the criteria for restricted verb-noun collocations were selected. Primarily, the pair had to consist of a verb and a noun. Concordance lines were reproduced and viewed for each pair to verify that the collocations were the verb-noun collocations. Secondly, the verb had to have a meaning other than its canonical meaning. In other words, the verb becomes delexicalized when co-selected with the noun in the pair; for example, the verb "make" in "make a decision" does not retain its primary meaning of creating or developing something.

After the automatic search and the manual selection process, the third step was to verify the raw frequency for each collocation. Using the same method for manual selection,

raw frequency counts were collected for various word forms for each of the 50 target collocations. For the target collocation "make a decision," for example, raw frequency counts were recorded for other variations such as "make decisions," "made a decision," "making a decision," "a decision was made," and other constituency and positional variations that were identified in the automatic search, producing a joint frequency count for each collocation. Table 3.1 shows an example of raw frequency counts for constituency variations for the collocation "make ~ decision."

Table 3.1. Raw joint frequency counts for constituency variations for the collocation "make ~ decision"

Collocate	Node	Joint frequency	Total joint frequency
Make	Decision	145	1175
Made	Decision	160	
Making	Decision	382	
Make	Decisions	186	
Made	Decisions	193	
Making	Decisions	109	

Concordance lines for the most frequent collocations were then used to develop the items on the test. This process was completed for 50 items in order to provide a large enough sample to develop a measurement instrument based on high-frequency collocations in written academic discourse. The development of a large number of items was intentional so that some items that did not perform well could be removed to produce a final test with fewer items.

A focus on the most frequent collocations is one way to identify a reasonable number of collocations systematically, relevant to the particular target language domain that test-takers are likely to know based on their exposure to academic English. The English language,

however, has a large number of frequent single lexical items but a small number of frequent multiword collocations.

Collocations were therefore not selected based on the frequency of individual word but on the frequency of a collocation as a single lexical unit in order to identify the true frequency of a collocation as it appears in a real-life target domain. This is not to say that the words which make up a collocation in a target language domain do contain all highly frequent words or words that are frequent in that particular domain. Studies have indicated the difficulty with measuring knowledge of collocations if the individual words that make up the collocation are not known (e.g., Schmitt, 1998), and a number of studies have tried to identify collocations with frequent constituents by selecting the individual items to create the collocation pair (e.g., Gyllstad, 2005, 2009). These studies, however, have sampled from an unknown English domain rather than a specific domain. Without specifying a domain, the potential number of collocations may increase, but the likelihood of a test-taker knowing all constituents in any particular collocation may decrease. A frequency-based approach to the identification of collocations as a whole unit thus will be a more precise indication of their frequency of the word pair that makes up the collocation in a specific target language domain.

Given the nature of the written academic collocations that were sampled from a specific domain, the participants in higher proficiency levels with higher scores on both the vocabulary levels test and the word association test would be expected to perform better on the collocation test. Ideally, this will also lead to positive washback helping students identify appropriate collocations and study more efficiently. Although evidence of positive washback

would support intended score use, the test would need to be operational and additional research necessary for this backing to be presented in this paper.

3.1.3 Task Design

The test task was developed based on task characteristics and conceptual framework described in chapter 2. A sentence-level gap-filling task held the most promise for eliciting collocational ability. This task is a favored approach to collocation instruction in the classroom (Coxhead, 2008; Webb & Kagimoto, 2009); therefore, a test with a similar task would be appropriate so that collocations might be tested the same way they are taught. This task has additional benefits. First, it discourages strategies that the test-taker would use to avoid production of collocation pairs in a free production task. Second, it contextualizes the collocations so that they are presented to the test-taker in an academic context.

Initially, 50 items were developed using the sampling process described in more detail below. The 50 items are listed in Appendix B and C. Sentences for the gap-filling items were chosen from the concordance lines in the general written academic sub-corpus for each target collocation in order to provide the context. An example of the concordance output can be seen in Appendix D.

For the pilot test, the 50 test items were contextualized in authentic academic language sampled from the academic sub-corpus of the BNC. The collocation form with the most frequent constituency and positional variation for each collocation was used to develop each item. Sentences were selected from the concordance lines from each high-frequency variation. The shortest concordance line that contained sufficient context was chosen. If no concordance line was shorter than 50 words, a longer line was truncated to fit the item

specifications. This was done by deleting phrases that contained additional but unnecessary information.

The declarative and interrogative sentences ranged from 10 words to 49 words including the gaps which were to be filled in with an appropriate collocate. The mean sentence length was 22.31. Most of the words on the test were highly-frequent. The majority (84.65%) of the total 777 running words appeared in the 1000 and 2000 frequency band of the General Service List (GSL), providing support for a claim that high-frequency collocations that are considered single lexical items are frequent enough to contain individual constituents with high-frequency.

Which constituent is the node and which is the collocate in a collocation is not always clear. For this test of collocational ability, the verb was chosen as the collocate. The most obvious reason for this choice was that the verb is the constituent that most often becomes delexicalized. A second reason that stems from research shows that L2 learners have more trouble selecting and/or producing the verb in a verb-noun word combination. Nesselhauf (2003), for example, identified 213 classified verb-noun collocations as restricted collocations in 32 essays. Of those classified as incorrect combinations, the most frequent mistake was a wrong choice of verb (p. 231).

The following two examples show the item type on the test with the verb replaced by a gap in the sentence. In Example Item 1, the test-taker is asked to fill in the verb that best completes the meaning of the sentence for an academic text. A verb in its base form was missing in this example. Other test items elicit the missing verb in other tenses or forms. The gap in Example Item 2 has replaced the verb in the past tense.

Example Item 1: It is difficult to _____a decision when you have two good choices. [key: make]

Example Item 2: A distinction can be _____ between planned and unplanned decentralization. [key: made]

3.2 Development relevant to the evaluation inference

The evaluation inference is based on the warrant that observations of performance on the collocation test items are evaluated to provide observed scores that reflect levels of collocational ability. This section describes the two scoring methods used in the study.

The dichotomous scoring method for this test was based on the target responses that were identified in the collocation selection process. The automatic corpus interrogation produced a list of the most frequent verb-noun collocations in the academic corpus. The collocations were identified based on frequency of both constituents, which identifies a target collocate for each node together as a whole unit. Although other collocates may be possible, they have a different frequency in the corpus and when they co-occur with the node, may not be semantically appropriate in the same context. Credit was awarded to a response that was based on the target collocation as identified in the sampling procedure. All other responses were marked as incorrect. This procedure is similar to Revier's (2009) study, which applied dichotomous scoring based on collocates identified through the frequency of whole collocations in the target corpus.

Awarding full credit for responses using the polytomous scoring method was similar to the dichotomous method. Responses that matched the collocate co-occurring in the target

collocations that were sampled from the academic sub-corpus of the BNC were awarded full credit of two points. Other responses were verified in the 86-million-word academic sub-corpus of the Corpus of Contemporary American English (COCA) (Davies, 2008). If a response was found to appear five or more times in the COCA sub-corpus as a collocate of the node in the item, one point was awarded as partial credit. No points were awarded to a response that was found less than five times in the COCA sub-corpus or not at all. A summary of both scoring methods is shown in Table 3.2.

Table 3.2. Summary of both scoring methods

Dichotomous scoring	Polytomous scoring
Correct: (1 pt) most frequent collocate from 16-million-word academic BNC sub-corpus (e.g., made, make, makes, maken, making) Incorrect: (0 pt) all other responses (e.g., consider, created, provided, found, known, identified, selected, chosen)	Full Credit: (2 pts) most frequent collocate from 16- million-word academic BNC sub-corpus (e.g., made, make, makes, maken, making) Partial Credit: (1 pt) responses found with ≥5 hits in 86-million-word academic files of COCA (e.g., consider, created, provided, found) Incorrect: (0 pt) responses not found in COCA and responses in COCA with fewer than 5 hits (e.g., known, identified, selected, chosen) COCA = Corpus of Contemporary English
Max score: k = 35	Max score: k = 70

3.3 Development relevant to generalization inference

This section describes the test development issues that were pertinent to the generalization inference which rests on the warrant that expected scores are comparable across the items on the test.

Because the generalization inference requires an analysis of consistency, the pilot study required a sample of test-takers who would be representative of the target test-taking population. Participating in the pilot study were 13 test-takers from three proficiency levels: one group of native speakers (n=4), one group of teaching assistants (n=4), and one group of non-English majors at a university in China (n=5). The results from this sample guided the process of reducing the number of test items from 50 to 35 while keeping the reliability estimate consistent. The test was piloted with non-native test-takers and native speakers of English at the university. Based on the results of the pilot study, poorly performing items were eliminated to produce a test of 35 items with better item discrimination while maintaining a high reliability estimate. The items were eliminated based on the results from item-test statistics obtained from the pilot results. Fifteen items were discarded based on the following procedure. First, the SPSS reliability analysis indicated that removal of 10 items would result in a slightly higher reliability estimate; doing so increased the reliability from $r = .93$ to $r = .94$ for the NNS data set (N=69). These were items 4, 8, 9, 13, 16, 36, 37, 40, 43, 46, and 47. Secondly, five items that received no correct target response by native speaking test-takers were eliminated. Two items (#46 and #47) had already been discarded due to the reliability analysis. After eliminating the other three items (#20, #34, and #50), the alpha reliability did not change but remained high at $\alpha = .94$; however, a final item (#14) was

eliminated, keeping the alpha reliability at $\alpha = .94$ for 35 items. After eliminating the 15 items, none of the remaining items has an item-total correlation of less than .30. After discarding items 4, 8, 9, 13, 14, 16, 20, 34, 36, 37, 40, 43, 46, 47, 50, thirty-five items remained with an acceptable reliability estimate.

3.4 Development relevant to the explanation inference

The explanation inference is based on the warrant that the construct of collocational ability as assessed by the Collocation Ability Test accounts for the academic collocational language performance in academic discourse in English-medium colleges and universities. A key issue in the construct definition and sampling procedure described earlier is the notion of frequency as it relates to the potential difficulty of items on the test. This section presents a discussion of the theoretical backing that led to the decision to identify verb-noun collocations by frequency. The section begins with a description of the frequency of all vocabulary on the Collocational Ability Test. Also discussed is the notion of collocation as a whole unit.

Assuming that words with higher frequency are learned earlier than words with other frequencies and that this applies to multiword lexical items, then the assumption can be made that higher frequency collocations would be known by more test-takers. Consequently, the more frequent collocations from the corpus could be assumed to have a higher item mean (item facility [IF]) for learners who are more familiar with academic discourse. The collocations from the BNC academic sub-corpus were grouped into frequency bands based on raw frequency counts.

The total number of tokens and types for all test items including nodes and collocates are included in Tables 3.3 and 3.4. Overall, 83.65% of all running words (tokens) in the 35 test items come from the 1000 and 2000 frequency bands of the GSL as well as 77.08% of all word pairs. The collocates consist of 80% unique words (types) in the top 2000 most frequent words (target responses on the test), and nodes are made up of 75.76% word types in the top 2000 frequency bands.

Table 3.3. Total number of tokens and types for all test items (k = 35)

Frequency band	Tokens	(%)	Types	(%)
1000	605	77.86	254	61.35
2000	45	5.79	40	9.66
3000	65	8.37	60	14.49
Off-list	62	7.98	60	14.49
Totals	777	100%	414	100%

Table 3.4. Total number of tokens and types for all word pairs, collocates, and nodes (k = 35)

	Word pairs (concgrams)		Verbs (collocates)		Nouns (nodes)	
Frequency band	Number of types (content)	Percentage (%)	Number of types (content)	Percentage (%)	Number of types (content)	Percentage (%)
1000	31	64.58	10	66.67	21	63.64
2000	6	12.50	2	13.33	4	12.12
3000	9	18.75	2	13.33	7	21.21
Off-List	2	4.17	1	6.67	1	3.03
Total	48	100%	15	100%	33	100%

A clear distinction is apparent between the frequencies of constituents in the collocation pairs and frequency of collocations as single lexical items. The collocates in the group with the highest IF values all come from the 1000 most-frequent word band, whereas only 66.67% come from this band in the items with lower IF values. A similar trend can be seen with the nodes. Eighty percent of the nodes from the most frequent word pairs in the items with high IF values come from the 1000 frequency band. The nodes in the items with lower IF values consist of 55.56% in the 1000 band. Overall, it appears that the frequency of the individual constituents in the target word pairs is related to the overall frequency of the word pair as a single lexical item.

Theory suggests that knowledge of a word is found along a continuum in relationship to other words in the learner's lexicon (Meara, 1996, 1997), similar to the cline between free combinations that are produced following a rule-based selection and the idiom principle, which are more formulaic and fixed in form (Sinclair, 1991). The notion of groups of words stored as individual units in the mind has been an area of research for quite some time and is not uncommon in research on lexical bundles, formulaic sequences, formulaic language, prefabs, or multiword items. The notion of a multiple-word unit as a single holistic lexical unit will provide the basis for understanding lexical processing and has implications for the unit of acquisition in second language vocabulary acquisition research and the construct definition in second language vocabulary assessment.

The concept of word strings as single units in the mind is the basis for the psycholinguistic study of formulaic sequences and collocations. Research has been conducted to investigate the psycholinguistic reality of word combinations stored as holistic units in L1 research (Durrant, 2008; Durrant & Doherty, 2010) and L2 research (Schmitt,

Grandage, & Adolphs, 2004; Siyanova & Schmitt, 2008; Wolter & Gyllstad, 2011). These studies sought evidence of relationships between lexical items or multiword lexical units stored in the mind.

The typical method of measuring the reality of word combinations in the mind has been to use a priming experiment, usually a lexical decision task (LDT) with software. *Priming* suggests that the presence of one word (the *prime*) activates a connection with a target word in the mind presenting evidence of a psychological association (Hoey, 2005). In such experimental situations, participants are presented with two words in sequence. Reaction times are measured when the second word appears until the participant indicates, with a keyboard stroke, recognition of the pair as a real word combination or not. Experiments with L1 participants have presented results suggesting that certain word combinations are psycholinguistically real for a variety of collocation types, verb-noun (Wolter & Gyllstad, 2011), adjective-noun, and noun-noun (Durrant, 2008; Durrant & Doherty, 2010). Positive forward and bidirectional priming effects have been presented for L1 research results, suggesting that multiword lexical units are stored holistically by L1 speakers.

Similar research investigating the psychological reality of involving L2 participants was not as confirmatory. Research with L2 learners has included a variety of collocation types, verb-noun (Wolter & Gyllstad, 2011) and recurrent strings (Schmitt, Grandage, & Adolphs, 2004). Not all procedures for research on L2 learners included reaction time as a time measurement, as in the L1 processing studies. Processing time was accounted for by the spaced repetition between utterances in the dictation task (Schmitt, Grandage, & Adolphs, 2004). Evidence of formulaic strings stored in memory was the accurate reproduction of the

strings in the dictation. Utterances were long enough to discourage short term memorization. The authors conclude that non-natives have very few formulaic sequences stored in their minds that are available for fluent and accurate language production.

Among other studies using a similar methodology as the research involving L1 participants, Wolter and Gyllstad (2011) conducted a primed lexical decision task with L1 and L2 learners using three conditions for verb-noun collocations, collocations with L1 equivalents, collocations with no L1 equivalent, and word pairs with little or no relationship. Initial findings indicated significant differences in all priming effects for NNS. However, in the condition where the collocation had no equivalent in the L1, only some of the items responded to priming whereas others did not. This finding is promising for evidence of multiword processing in a second language.

In summary, it appears as though reaction times indicate that native speakers can recognize multiword units faster than they recognize word combinations that are not considered strong collocations. This finding is considered to be evidence of storage of such sequences in the mind that can be quickly accessed and produced as a contributing factor in fluency. Research with non-native speakers shows some indication of similar evidence but, overall, little indication of either knowledge of formulaic language or storage of formulaic language as multiword units. With the exception of the dictation task, a common theme for all of the research thus far is the consideration of successful recognition as evidence for the storage of collocations in the mind. L2 learners tend to approach language learning through a rule-based system that considers a single word the primary building block and the recognition of decontextualized word combinations. A discussion of vocabulary acquisition theory is beyond the scope of this study, but there is some evidence to indicate a tendency for

L2 learners to either store multiword units or strengthen the relationship between individual words to the point that reaction times and language production replicate acceptable use in a target language.

The basic underlying theory of the importance of knowing high-frequency words has been the fact that they cover a large proportion of written and spoken language in many contexts (Nation, 2001), which has led to the widely-accepted theory is that the more frequent a word occurs in a language, the more easily and likely it is to be learned (Milton, 2007). This idea has been applied to formulaic languages and collocations that include the phrase *frequently co-occur* in their definition. The notion of frequency has two related perspectives. *Frequency* can describe the words co-occurring together more frequently than by chance, but it can also describe the raw frequency of the collocation occurrence in language sample. The frequency of a word (or formulaic sequence) has been an important factor in selecting items for priming studies.

In light of the likelihood that test-takers may have encountered target collocations in academic English in their courses or English language learning, one might consider the relationship between *frequency* of a collocation and *item difficulty* of the item with the collocation on the test. The relationship between frequency and the psycholinguistics of collocations has been investigated using recognition tasks with native speakers. Frequency is based on analysis of corpus data that can represent a sample of language use for a particular genre or register or undefined collections of texts. Occurrences of collocation can be distributional and context specific, whereas other collocations can be found in many contexts. Initial inquiry into this topic considered a large sample of text to be most representative of language use. The BNC was the preferred source because of its accessibility and size.

Ellis, Frey, and Jalkanen (2009) concluded that higher-frequency verb-object and booster and adjective collocations are more readily perceived than collocations with lower-frequency of items identified in the BNC. Corpus-driven item selection for research on item frequency and priming indicate that higher-frequency formulaic language has significant priming effects for adjective-noun and noun-noun pairs (Durrant, 2009; Durrant & Doherty, 2010; Schmitt, Gradage, & Adolphs, 2004). One can intuit that the more frequently a word or word combination is used, the more well-known the words or combination is in the mind of the language user. Consequently, the more frequent a word or collocation, the easier and earlier it will be learned.

An overview of these studies seems to indicate a trend that frequency does play a role in facilitating the recognition of collocations; however, these studies developed frequency lists from corpus data and grouped the formulaic language by bands rather than providing explicit frequency information. Mean scores by group are presented which could confound the performance by combining very short and longer frequency ranges by category, revealing little about the difficult of individual items.

Research on collocational knowledge by L2 learners began sampling collocations from corpora based on frequency primarily due to the widely accepted definition of a collocation being composed of words that co-occur frequently. The approach followed either the word-based tradition by selecting frequent single words as nodes and then using software to identify frequent collocates in the corpus or the phraseological tradition by identifying collocations based on the frequency of both parts together. This could be done by letting the software identify the items (corpus-based) or selecting items and then verifying the frequency with the software (corpus-verified). The underlying theory for the use of highly frequent

words in a test of collocational knowledge is based on the accepted definition of collocation as frequently occurring word combinations but with multiple interpretations of what this means.

The frequency approach also fulfilled a second assumption that the relationship between words should be measured using words that are minimally known by the L2 learner (Gyllstad, 2009). Schmitt (1998) was confronted with this issue when the test-takers in his study were unfamiliar with the prompt words in the task.

Despite the fact that most researchers using corpus data follow some type of frequency-based sampling, the results presented only distinguish among proficiency groups and offer no relationship between frequency and item facility; thus, even significant effects between higher proficiency and lower proficiency groups on a collocation test may be evidence of overall language proficiency rather than a correlation with item frequency and facility. Eyckmans (2009) claimed significant differences among proficiency groups on a pre-test but not on a post-test following instruction. Scores on a collocation test may show less discrimination after treatment.

Durrant and Doherty (2010) report lower reaction times for high-frequency adjective-noun combinations; however, research on frequency as a dimension of difficulty in collocational knowledge for L2 learners has been inconclusive and counterintuitive. Using the terms *native-like* and *frequent* interchangeably as well as *non-native-like* and *atypical* in a judgment task of adjective-noun combinations, Siyanova and Schmitt (2008) found that native speakers were better at judging collocations by frequency, yet L2 learners rated infrequent collocations as more familiar and frequent collocations as less familiar than native speakers did.

The underlying finding is that more proficient L2 language users are more competent at recognizing and producing appropriate word combinations regardless of frequency, despite the claim that the frequency of encounters contributes to vocabulary learning more than contextual richness (Joe, 2010). In addition, exposure as a single factor has been viewed as positive in the acquisition of word combinations as well. Siyanova and Schmitt (2008) found evidence that more L2 exposure does lead to better intuitions about collocations. Numerous factors have influenced these studies, including word- or phrase-based identification of collocations and possibly even type of collocation. In sum, the connection between frequency and item difficulty is still under investigation.

3.5 Chapter summary

This chapter has described issues in the development of the Collocational Ability Test as they relate to some of the inferences in the interpretive argument for the test. The chapter presented and discussed issues related to the following inferences: domain description, evaluation, generalization, and extrapolation. No direct issues related to the extrapolation, utilization, or intended impact inferences were raised during the development of the test, other than the selection of participants. Evidence to support the utilization and intended impact inferences needs to be collected after the test is used for its intended purpose. The participants were selected to represent a variety of proficiency levels related to study at an English medium university. The next chapter on methodology provides a description of the design of the study, presenting information on participants, materials, measures, procedures, data collection, and data analysis.

Chapter 4 Methodology

This chapter presents the methodology used in this dissertation, which guided the collection of empirical evidence to support the warrants and claims in the interpretative argument for the test of collocational ability. This study followed a mixed-methods embedded and sequential explanatory design to collect data from 206 participants. The participants represent various levels of language proficiency and student classification in the intensive English program and matriculation as an undergraduate or graduate student at Iowa State University.

This chapter also presents the materials and instruments that were used in the study: the Collocational Ability Test, a reading test, a vocabulary size test consisting of three sub-tests, video from screen capturing during administration, and questions leading a post-test interview and a Test Reflection Survey. The chapter describes the procedures that were used in the study, beginning with an overview of the mixed-methods embedded and sequential explanatory design. The procedures for data collection and data analysis of the quantitative data are explained followed by discussion of the procedures for collecting and analyzing the qualitative data.

4.1 Design

This study was designed to collect evidence to support the chain of inferences in the interpretive argument of the Collocational Ability Test. Data collection was guided by the research questions, which were developed in order to support various assumptions underlying three of the inferences in the interpretive argument: the generalization inference, the explanation inference, and the extrapolation inference. The mixed-methods embedded and

sequential explanatory design in this study consisted of two phases of data collection (Creswell & Plano Clark, 2007). The study design called for a collection of both quantitative and qualitative data. The purpose of this design was to explain the results of the quantitative data with results from the supplementary qualitative data.

4.2 Participants

The 206 participants in this study were sampled from three potential stages of matriculation at Iowa State University. The participants were selected because they are a representative sample from a population of intended test-takers for the Collocational Ability Test. Proposed test-takers for this test are students who are entering an English-medium institution of higher education. The three groups of test-takers are shown in Table 4.1. The groups are classified by level of study, resulting in one group of graduate assistants, the high-ability group (n=35), one group of undergraduate university students, the mid-ability group (n=109), and one group of intensive English students, the low-ability group (n=62).

Table 4.1. Participants in the study by group

Group	N	Description
High-ability Group	35	Graduate assistants working as ISU RAs & TAs
Mid-ability Group	109	Undergraduate students enrolled in ENGL 99 & 101 courses
Low-ability Group	62	Intensive English Language students at IEOP
Total	206	

The participants in the first group, the low-ability group, are students in the Intensive English and Orientation Program (IEOP) at Iowa State University (ISU). IEOP is an

intensive English language program that prepares students for study at an English-medium university. The program offers language instruction and orientation activities that introduce the students to academic life on campus. The program has three bands and six proficiency levels, with instruction in reading, writing, grammar, and oral communication. Students are placed into classes based on their placement test scores. A diagnostic test is then administered to each class at each level, and students who perform well on the diagnostic test are considered for advancement to the next level. At this time, admission to the university is not linked to successful completion of the program. Many students are admitted to the university with scores from standardized tests before completing the program. Because of their placement into an intensive English program, this group is referred to as the *low-ability group*.

The second group of participants, the mid-ability group, consists of undergraduates who are enrolled in non-credit ESL courses that focus on developing academic language skills. Performance on the university's English placement test indicated that these students would benefit from additional English instruction before continuing in their discipline at the university. These language courses are completed within the first year. The participants in this group may have some experience with academic English but need more instruction and exposure to academic language at an English-medium university. This group is therefore referred to as the mid-ability group.

The third group, the high-ability group, is comprised of graduate assistants who are working as research or teaching assistants for a variety of departments at the university. These students are working as graduate assistants in addition to completing their research in their discipline. The 16 academic disciplines that are represented by the participants are:

Aerospace Engineering, Art and Design, Biochemistry/Biophysics, Chemistry, Computer Science, Curriculum and Instruction, Economics, Educational Leadership, Electrical and Computer Engineering, English, Genetics, Horticulture, Journalism and Mass Communication, Natural Resources Ecology, Plant Biology, and Veterinary Microbiology and Preventative Medicine. These participants are actively engaged in their discipline and frequently encounter the academic language that is used in their field of study at the university.

All participants in this study were Chinese-speaking learners of English. The selection of test-takers with a common first language was intended to limit the potential variability in responses that could be associated with L1 transfer.

4.3 Materials and instruments

This section describes the six data collection instruments, the Collocational Ability Test, the reading test, the vocabulary size test, the Test Reflection Survey, screen capturing observations, and questions for the post-interview. All three tests and the Test Reflection Survey were delivered via the Internet using the Moodle learning management system using Firefox browser. Each test-taker was authenticated with an individual username and password. All tests were administered in computer labs using iMacs running Mac OSX Version 10.6.8 on a 21-1/2" screen with a resolution of 1920x1080 with black text on a white background. The test items were not available to the test-takers outside of the testing period. No test could be taken more than one time. All tests were timed. Test-takers could finish early or use the entire time allowed.

4.3.1 Collocational Ability Test

The Collocation Ability Test is a 35-item computer-delivered test. The development of this test was described in the previous chapter. Each item is a sentence which contains one of the 35 target collocations that were identified though the corpus analysis. The verb, the collocate of the collocation, has been removed and replaced with a gap for each item on the test. The test-taker enters the missing collocate in a text field beneath the sentence shown on the screen. A screenshot preview of the collocation test is shown below in Figure 4.1. Sentence length ranges from 10 to 49 words with an average of 22 words. The instructions for the test appeared at the top of the screen.

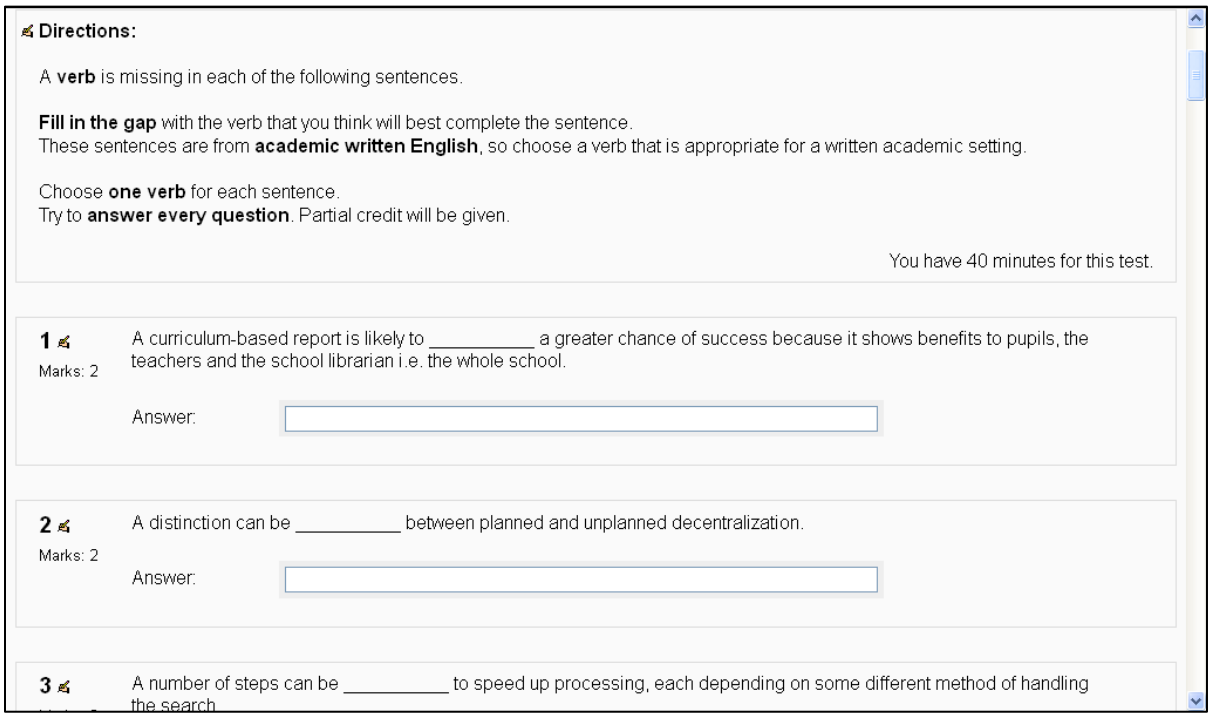

Figure 4.1. Screenshot of the Collocational Ability Test

The time allowed for the Collocational Ability Test was 40 minutes. A clock in the upper left corner of the screen counted down the time. If a test-taker completed the test before the allotted time, he/she could click the "Submit all and Finish" button at the bottom of the screen. If time ran out, the system would save all of the answers provided and close the test.

4.3.2 Reading test

The reading test was a 20-item multiple choice reading sub-test from the Michigan English Language Assessment Battery (MELAB) that measured the test-taker's reading ability. The test consisted of 20 short passages with a multiple-choice question format. Test-takers read passages and answered multiple-choice questions about the passages as an indication of their reading comprehension and placement into a reading class for the intensive English group. The test settings were configured so that the distractors were shuffled for each item to appear in a random order for each test-taker. Time allotted for this test was 20 minutes, with a clock counting down the time in the upper left corner of the screen. Similar to the Collocational Ability Test, a test-taker clicked on the button "Submit all and Finish" at the bottom of the screen to complete the test, unless time ran out, saving the responses that had been provided during the 20 minutes.

4.3.3 Vocabulary size test

The vocabulary size test was a 30-item test requiring production of word in gap-filling task. The gap contained the first few letters from the target response. Each item was presented as an individual sentence rather than multiple gaps in a coherent text. This format

was based on the productive levels test developed by Laufer and Nation (1999). Test-takers used the first few letters that were provided and completed the response in a text field on the screen to provide an appropriate target response. A screenshot preview of the vocabulary test is shown in Figure 4.2. Instructions for the vocabulary size test were provided for the test-takers at the top of the screen.

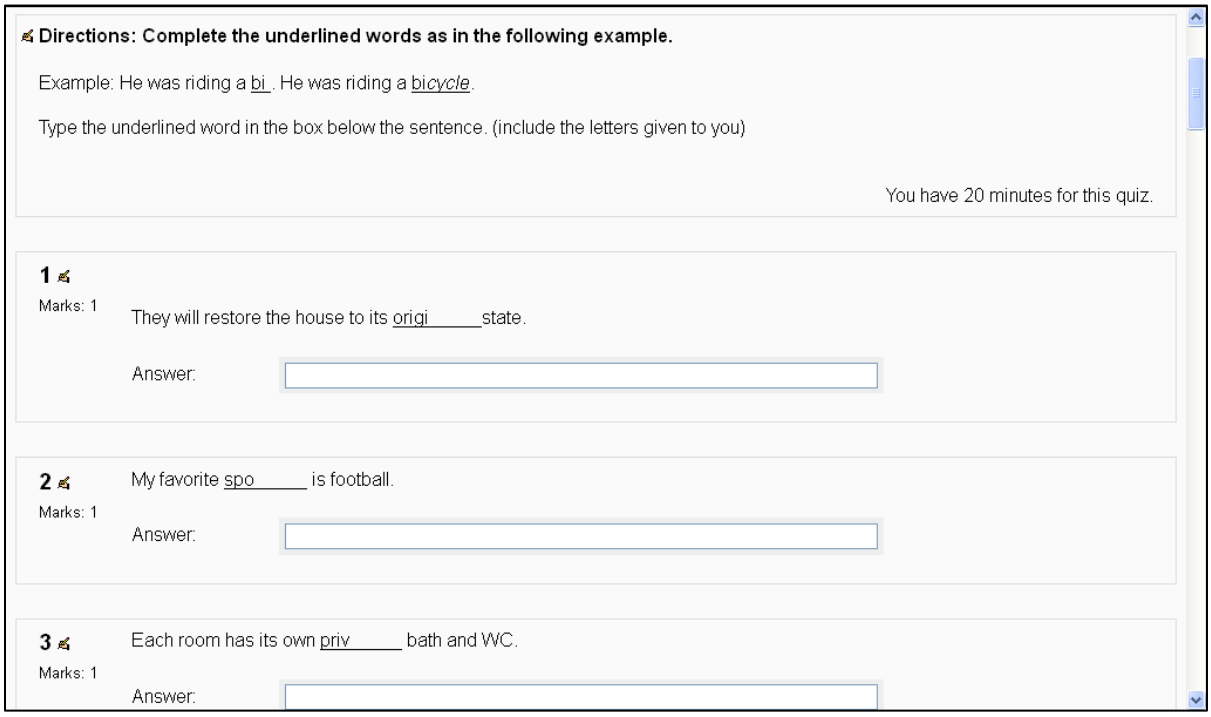

Figure 4.2. Screenshot of the productive vocabulary test

Each of the three 10-item sub-tests measured knowledge of a unique sample of single lexical items. The first sub-test measured knowledge of high-frequency general English vocabulary sampled from the GSL word list in the 2000-word frequency band. The second sub-test was constructed with words that made up the target collocations and content words

on the test. The third sub-test was developed using words from an academic word list (Coxhead, 2000). The target words on the vocabulary size test are shown in Table 4.2.

Table 4.2. Lists of target items on the vocabulary size sub-tests

2K Items	Content Items	Academic Items
Original	Attract	Affluence
Sport	Collect	Episodes
Private	Damage	Innovation
Total	Attention	Deficit
Elect	Information	Prestige
Manufacturing	Role	Configuration
Victory	Data	Discourse
Melt	Conclusions	Hypotheses
Hide	Obtain	Anonymous
Invite	Evidence	Indigenous

Test-takers were allowed 20 minutes to take the vocabulary test. Time was monitored by the clock counting down the time in the upper left corner of the screen. Similar to the other two tests, a test-taker clicked on the button "Submit all and Finish" at the bottom of the screen to complete the test, unless time ran out, saving the responses that had been provided during the 20 minutes.

3.4 Test Reflection Survey

Three questions following the Collocational Ability Test composed the Test Reflection Survey, which elicited responses from the test-takers regarding their awareness of engaging in the use of academic language during test administration and the relationship between academic language on the test and in textbooks used at the university. This instrument collected both quantitative and qualitative data. It included three questions in

English with equivalent Chinese translations. The English prompts were translated into Chinese by a Chinese speaker and confirmed by two other Chinese speakers. The first two questions on the survey were multiple-choice with three with three options: “Yes,” “No,” or “I don’t know.” The third question was an open-ended response format with a text field for the response. The three Test Reflection Survey questions with Chinese translations were: Were you thinking about academic English as you took the test?

请问你在做该考试题的时候，同时有想到学术英语吗？

Do you think the English in this test is similar to academic English used in university textbooks?

请问你认为该考试的题目与大学课本中的学术英语的应用相似吗？

Explain how the English in this test is similar to or different from English used in university textbooks.

请解释一下该考试的题目与大学课本中学术英语应用的相似性或者不同性。

4.3.5 Screen capturing observations

The video feature of Camtasia 1.2.2 software by TechSmith was used to capture the screen during test administration for a sample of test-takers. The software interface disappeared after it had been started, and the only indication of recording was a small red light at the top of the browser. Only the screen was captured; the camera was not used to record the test-taker viewing the screen. Audio recording was also used during the administration, but because test administration was fairly quiet, no audio contributed to the data interpretation. The software was started before the test began. Test-takers were aware

that the screen was being recorded. A sample screen shot is shown in Figure 4.5. The video files were compressed and saved in MP4 format for later analysis. The file size ranged from 43 to 175.3 MB.

4.3.6 Questions for semi-structured post-interview

A semi-structured follow-up interview was conducted for a sample of test-takers. The purpose of the interview was to investigate the meta-cognitive strategies that the test-takers employed during the test. Test-takers were interviewed individually with the following script to prompt the discussion. The structure of the semi-structured interview is listed below.

Opening question:

- Did you understand what the test was testing?

Discussion:

- Were you thinking about academic English as you took the test?
- Did you have difficulty understanding any of the sentences?

Further discussion (construct):

- Were you taught or did you study verb-noun collocations in class or English study?
- Did you memorize word-pair lists or did you study collocations in context?
- Questions about the first-language transfer for relevant test items.

4.4 Procedure

The first step was to obtain approval from the ISU Institutional Review Board prior to solicitation of participants and data collection. Volunteers were solicited by email and

personally in class by the researcher at Iowa State University in the United States. The Collocational Ability Test, reading test, and vocabulary size test were administered to the volunteers over a 3-month period during the third quarter of 2011, in a computer lab on campus, in 1-hour sessions. The study followed a followed a mixed methods embedded and sequential explanatory design consisting of two phases, as presented in Figure 4.3.

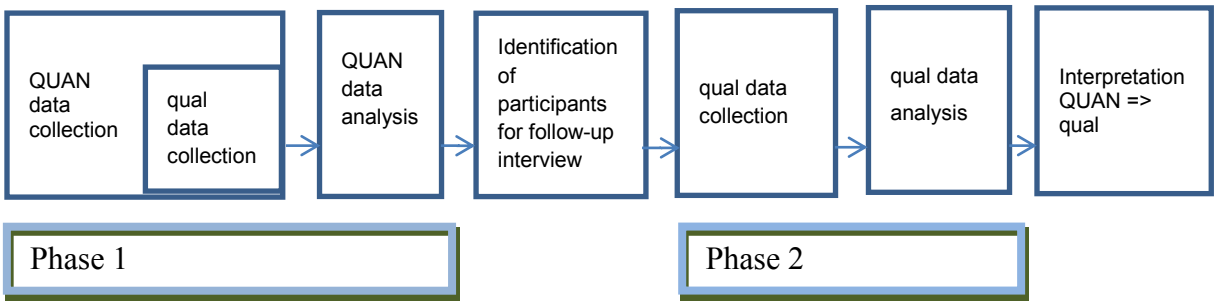

Figure 4.3. Overview of the mixed-methods research design used in this study

4.4.1 Phase 1: Quantitative and embedded qualitative data collection

Quantitative data was collected during Phase 1 during the beginning of the Fall 2011 semester using the Collocational Ability Test, reading test, and vocabulary size test. The embedded aspect of the research design accounted for qualitative data to be collected from the Test Reflection Survey for all test-takers and screen capturing for a sample of test-takers during the initial phase when the quantitative data was being collected.

The low-ability participants took all tests along with other placement instruments for the intensive English program. All tests were administered to the mid-ability participants as part of a class assignment during two class periods. Because the tests were administered as part of an English language class, instruction about collocations followed the test administration in exchange for the class time taken by the test administration. Although all

students in the classes took the tests, instructors for the courses identified the Chinese-speaking students in this group so that only those students' test scores would be used in this study. Motivation to complete the tests was high, because the administration took place during class time. The researcher provided instructions and, along with the classroom teacher, monitored the test-takers during the test administrations to prevent activities such as talking, note-taking, and browsing other websites during the test. The high-ability test-takers took the test at different times throughout the data collection period. This group was purpose-driven and focused on the tasks without any difficulty. Motivation and concentration did not appear to be an issue for test-takers in this group. All test-takers took the Test Reflection Survey at the same time as the Collocational Ability Test. This instrument collected and stored responses in the learning management system (LMS).

A number of test-takers from the low- and mid-ability groups took the test at various times along with a number of high-ability test-takers as part of the embedded research design. These students from these groups took the tests with high-ability participants because they either arrived late for the beginning of the school semester or they missed the class when the test was administered. These test-takers, along with a sample of high-ability test-takers, formed the screen capturing group. They agreed to allow the screen to be recorded while they took the tests. Using Camtasia 1.2.2 software, the screen was recorded for a total sample of eight participants: four participants from the high-ability group, two test-takers from the mid-ability group, and two test-takers from the low-ability group. Video files were compressed after the test administration for storage and analysis.

The next section describes the second part of phase 1, which is the analysis of the quantitative data. Based on the results of the data analysis in phase 1, a number of test-takers were then selected for participation in the qualitative data collection in phase 2.

4.4.2 Phase 1: Quantitative data analysis

Initially, the results for the Collocational Ability Test, reading test, and vocabulary test were automatically scored by the LMS. The quiz module in LMS, Moodle, automatically scored the responses according to target responses, which had been manually entered in the answer key in the LMS. Initially, all tests were scored using a dichotomous scoring method.

The dichotomous scoring scale awarded 1 point to the response that matched the target collocate in the item that had been selected from the corpus-based sampling procedure. Variations in spelling, tense, and inflection were not included in the criteria for correctness. Responses that did not correspond with the target collocate received no credit. After test administration, the answer key was updated with alternate spelling or word forms that were identified in the responses and automatically regraded. Two files were then downloaded from the LMS for the Collocational Ability Test and the Vocabulary Test. The first file contained the scores from the tests using a dichotomous scoring method. The second file contained the response for all test-takers for all items. Only the dichotomous scores were exported for the reading test, because it was a multiple-choice format. The data were downloaded in Excel files and stored on a computer for analysis. The test-takers' names were replaced in the files with identification numbers.

The next step was to rescore the Collocational Ability Test using a polytomous scoring method, which began with a response analysis. Similar to the dichotomous scoring

method, the polytomous scoring scale considered a full-credit response as a match with the target collocate. This match was awarded two points. Partial credit was given to responses that were identified as academic commutable collates of the node in the item. Response analysis was conducted to identify responses that should receive partial credit.

Each response was verified using COCA's 86-million-word academic files (Davies, 2008). First, collocates for each item were identified by frequency, using the academic genre in COCA with a span of 4 to the left. The lists were truncated to contain all collocates with five or more occurrences. The test-takers responses were then compared with the list. Any test-taker response that was also on the list generated from COCA received partial credit, 1 point. This process was repeated for all 35 items. A response was acceptable in any form regardless of spelling, tense, or inflection. All other response were marked as incorrect and receive no credit, "0." This approach suggests that highly frequent collocations that are produced in a specific context may have partial association by the test-taker.

Table 4.3 shows test-takers' responses that were awarded full or partial credit for the item, "A distinction can be _____ between planned and unplanned decentralization," and the credit awarded to each response. The target collocate identified by the first corpus interrogation was "make" which was realized as "made" in the actual test item. A full list of all collocates identified in the academic files on COCA along with their frequency, all test-taker responses, and credit for this item are shown in Appendix E.

Forms of the target collocate "make" were awarded full credit of two points. The other collocates which were found in the list of collocates identified in the COCA search were awarded partial credit, one point. Responses are shown in Table 4.3 and Appendix E without correcting spelling errors. The dichotomous scoring methods resulted in four correct

responses. The polytomous scoring method resulted in four responses with full credit and an additional 17 responses receiving partial credit. The scoring scale based on this type of sampling procedure provides evidence in support of the evaluation inference in the interpretive argument. Data was then analyzed to provide backing for the generalization inference with reliability estimates and explanation inference through correlations with other tests.

Table 4.3. Test-taker responses and credit for the item for "make ~ distinction"

Test-taker responses	Credit
made	2
make	2
maken	2
making	2
consider	1
created	1
DO	1
done	1
finded	1
found	1
have	1
provided	1
see	1
seen	1
showed	1
showen	1
shown	1
understand	1
understood	1
use	1
used	1

Descriptive statistics and reliability estimates were calculated for each group for each test using Classical Test Theory (CTT) and were analyzed to check for normal distributions.

Histograms were presented to show the score distributions. The Collocational Ability Test was analyzed once using the dichotomous scoring method and a second time after being rescored using the criteria from the polytomous scoring method. Parametric ANOVA tests were used to identify statistical significance among groups.

Relationships among all measures and groups were calculated including both the dichotomous and polytomous results for the Collocational Ability Test. Correlational relationships were calculated between total scores on the Collocational Ability Test, and total scores on the reading and vocabulary tests were calculated using Pearson's correlations and Spearman's Rho correlations. Correlations were also calculated between the means of each test item and the frequency of each collocation. Scatterplots display the relationships visually.

An analysis using Item Response Theory analysis (IRT) was conducted on the sets of scores from the collocational Ability Test. The chosen IRT model for this study is the Rasch model with one characteristic of measurement, *item difficulty*. The item difficulty results indicate an alternative analysis to the correlation between *item facility* and *item frequency*. The advantage of Item Response Theory over CTT is to use mathematical models to predict probability estimates for both test-takers and items that are independent of a particular test administration or set of test items. The model hypothesizes an unobservable, psychological variable that underlies performance. The underlying trait or ability in this study is presumed to be academic verb-noun collocational ability. The hypothesized relationship between performance on the test and the underlying ability is applied to the IRT model to see if the data fit the model. When items "fit the model," they are assumed to be "sample free" which means they do not depend on a particular set of items on the test (Lynch, 2003). Results from

the Rasch analysis were presented in two ways, once the misfitting items had been identified. First, aggregate item characteristics curves were presented for both scoring methods. Then individual item characteristic curves were shown for each misfitting item in each scoring method.

The data from the Test Reflection Survey indicated the test-takers' perception of academic language on the test and in university textbooks. The data were analyzed to indicate whether the perceptions as groups were significantly different. The observed data from the first two questions on the Test Reflection Survey were collected in a contingency table using the three proficiency groups and the three potential responses: "yes," "I don't know," and "no." The data were then compared using Pearson's chi-squared statistic hypothesis test.

The open-ended question in the Test Reflection Survey prompted test-takers to explain how the English in the test was similar or different from English used in university textbooks. The responses were organized by test-taker group and analyzed from a grounded theoretical perspective to allow trends in each group to emerge. Results were downloaded to an Excel file for manual analysis. The responses from the third question on the Test Reflection Survey were analyzed by group—high-ability, mid-ability, and low-ability—for emerging themes. Using a grounded approach, the responses from these three groups were analyzed to identify familiarity with academic language. Differences were quantified and presented as percentages in a bar chart.

The Camtasia 1.2.2 video recordings from the screen capturing were saved as MP4 files and analyzed for evidence of how test-takers approach the tasks on the Collocational Ability Test. Evidence of metacognitive strategies was identified by a sequence of response

to an item by a test-taker. Most test-takers provided a response and went on to the next item or made small changes to spelling or verb tense. These changes were not recorded. Changes in responses were considered evidence if a test-taker provided a response, deleted it, and entered a different verb. Providing a series of responses to a test item was seen as an indication that the test-taker was consciously evaluating the appropriate word for the item. These changes were recorded for each group, quantified, and displayed as percentages in a chart.

4.4.3 Identification of participants for qualitative data collection

Following the initial collection of quantitative data, test-takers were recruited from the three identified groups in the study for follow-up interviews. In order to obtain a proportionate representation, test-takers were identified by group and dichotomous test score on the Collocational Ability Test. One student from each group with a high score for their group and one student with a low score for their group were recruited via email. A total of six test-takers were interviewed. Interviews were conducted individually and the audio was recorded or notes were taken during the interview if the participant did not allow audio recording. When recording was permitted, the audio was captured on a laptop computer using Audacity audio recording software. Table 4.4 shows the test-takers in each group who participated in the post-interview, along with their test scores for both scoring methods. These data show that the distinction between high-ability and low-ability test-takers remained for both scoring methods.

Table 4.4. Test-takers who participated in the post-interview

ID#	Group	Dichotomous collocation test score	Polytomous collocation test score
569	GA	18	24.5
1094	GA	3	14
940	UG	13	20.5
700	UG	2	9.5
825	IE	11	16.5
891	IE	2	7.5

4.4.4 Phase 2: Qualitative data collection and analysis

Qualitative data was collected for a sample of test-takers. Collection and analysis of the qualitative data was intended to identify evidence to explain initial quantitative results by including evidence of test-taking processes to help explain the statistical results.

The qualitative data came from screen capturing for a sample of test-takers (n=8) and semi-structured post-interviews conducted with three test-takers from each group (n=6)**.** The screen capturing during test administration for a sample or test-takers occurred simultaneously during the quantitative data collection in phase 1. The interviews were conducted in the weeks following the test administration.

A semi-structured post-interview was conducted with two test-takers from each group. Recruitment was based on their score on the Collocational Ability Test using the dichotomous scoring scale; one test-taker from each group with a high-score and one test-taker from each group with a low-score were solicited. An email was sent to the test-taker requesting a time for a follow-up interview. A test-taker who did not respond was not pursued. An alternate test-taker was then solicited. Six test-takers were interviewed individually in sessions lasting from 30 to 45 minutes. The interview included a review of

selected responses on the test. Interviewees were asked about the test content, previous experience learning collocations, and potential L1 transfer. Notes were taken for each interview, and audio was recorded using Audacity software when permitted by the interviewee.

The responses from the test-takers who participated in the post-interview were compared for similar trends among test-takers with similar performance on the test regardless of group membership. The exploration was seeking answers to the high or low performance on the test. The interview was semi-structured so that a discussion could emerge from comments made by the interviewee. Common responses were presented either in support of or against the results from the analysis of the quantitative data.

4.5 Chapter summary

This chapter explained the selection and elicitation of participants in the study. The materials and instruments were then described. The procedures used in the study were then detailed, starting with an overview of the mixed-methods design. The discussion of methodology included an explanation of data collection and analyses procedures. Procedures were explained following the design of the study, beginning with collection and analysis of quantitative data. The scoring method for the Collocational Ability Test was detailed, followed by the procedure for selection of test-takers for qualitative data collection. Finally, collection and analysis of qualitative data were explained. The next chapter introduces the results, following the research questions presented in chapter 2.

Chapter 5. Results and discussion

This chapter presents the results from the data collected to support the interpretive argument. The findings are reported as they address the research questions associated with each assumption that was identified in chapter 2. The results provide backing for each warrant, which supports to some degree each intermediate claim along the chain of inferences in the validity argument. Because one objective was to compare the quality of the validity argument under the conditions of two scoring methods, findings are presented to compare the dichotomous and polytomous scoring methods. The results of the two scoring methods are compared as an evaluation the efficacy of each scoring method and the degree of backing each provides to the validity argument. This chapter begins with a presentation of the descriptive statistics and histograms of the total scores on the Collocational Ability test for the dichotomous and polytomous scoring methods. It then presents empirical results that answer each of the research questions intended to provide additional evidentiary support for or challenge the assumptions underlying the inferences in the interpretive argument for the Collocational Ability Test. Support for the domain definition inference and part of the explanation inference are provided in the framework above. Empirical results addressed for each of the research questions provide additional backing for the assumptions underlying the evaluation inference, generalization inference, and extrapolation inference.

5.1 Descriptive statistics and reliability estimates

Descriptive statistics and reliability estimates were calculated to describe the performance of the each group of test-takers on the Collocational Ability Test using both scoring methods. These statistics were calculated to describe and compare the characteristics

of the groups and the consistency of scoring. The descriptive statistics also guided the selection of subsequent statistics for further analyses. Histograms displayed the distributions of scores visually. The descriptive statistics were useful in supporting the evaluation inference in the interpretive argument. Backing for the assumption that scores had appropriate characteristics for norm-referenced decisions was supported or challenged based on these statistical results.

5.1.1 Descriptive statistics and reliability: dichotomous method

Descriptive statistics for the scores on the Collocational Ability Test using the dichotomous scoring scale are shown in Table 5.1. The dichotomous scoring method awarded 1 point to the target response which had been identified by the corpus search. All other responses were marked as incorrect with 0 points. The mean scores decreased, as expected, from the highest placement group to the lowest placement group. The high-ability group performed the best (M=12.31), followed by the mid-ability group (M=5.01), and then the low-ability group (M=2.81). The reliability estimate calculated for the whole sample (N=206) was acceptable at $\alpha = .83$; however, the distribution of the scores is skewed, as shown in Table. 5.1

Table 5.1. Descriptive statistics for the Collocational Ability Test for all levels (k=35) using a dichotomous scoring method

Group	N	Min	Max	Mean	Std. deviation	Cronbach's alpha
High-ability	35	3	21	12.31	4.15	
Mid-ability	109	0	15	5.01	3.38	.83
Low-ability	62	0	11	2.81	2.75	
TOTAL	206	0	21	5.59	4.64	

Figure 5.1 shows a histogram of the distribution of scores for the Collocational Ability Test using the dichotomous scoring method. Although the initial values for skewness (0.957) and kurtosis (0.53) are within the "rule of thumb" values (-2 and +2) for a normal distribution (Bachman, 2004, p. 74), these statistics, divided by their standard of error, provide a picture of a non-normal distribution with the skewness/SES (5.59) over the maximum value and kurtosis/SES (1.75) reaching the maximum value. The histogram of scores confirms the idea of an asymmetrical non-normal distribution.

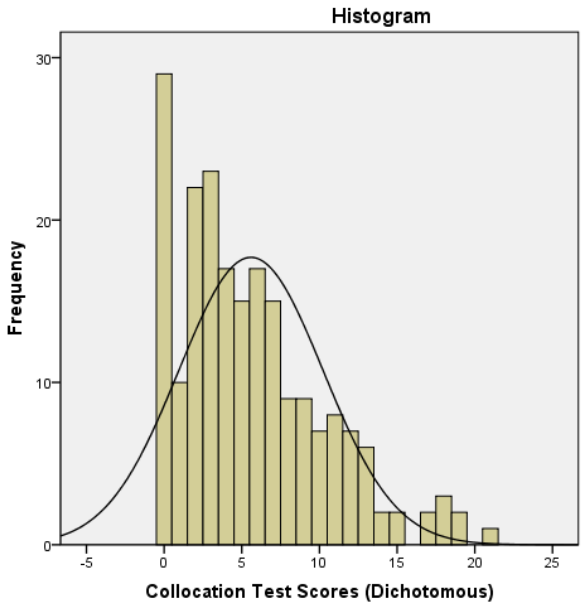

Figure 5.1. Histogram showing the distribution of scores on collocation test (k=35) for all groups (N=206)

A distribution with a positive skew as shown in Figure 5.1 indicates that the test is too difficult for the population represented by the sample of test-takers. This is the type of distribution that would be expected from the scores on a pre-test intended for a criterion-referenced interpretation. The asymmetrical distribution of scores confirms the information provided by the values from the descriptive statistics indicating a non-normal distribution.

Backing was therefore not found to support the assumption in the evaluation inference that that scores had appropriate characteristics for norm-referenced decisions could be supported or challenged based on these statistical results.

5.1.2 Descriptive statistics and reliability: Polytomous method

The data were recalculated by applying a partial credit scale to the results following the identification of acceptable relevant collocates in the corpus of written academic English. The polytomous scale allowed for partial credit if a response was identified as a potential collocate of the node in the item in the context of academic English. Response analysis, described in chapter 3, identified the potential collocates by frequency in the academic texts of COCA. A response that matched the target response was given full credit as 2 points. Responses that were identified as potential collocates in academic English were awarded one point. All other responses received no credit. The descriptive statistics and reliability estimate for the results using the partial credit scale are presented in Table 5.2.

Table 5.2. Descriptive statistics for the Collocational Ability Test for all levels (k=35) using a polytomous scoring method

Groups	N	Min	Max	Mean	Std. deviation	Cronbach's alpha
High-ability	35	21	52	36.91	7.18	
Mid-ability	109	0	41	22.67	8.82	.89
Low-ability	62	0	33	16.05	9.27	
TOTAL	206	0	52	11.54	5.54	

A similar trend is seen with the partial credit scale as seen with a dichotomous scale. Mean scores for all three groups descend as expected from the highest to the lowest proficiency group. The high-ability group performed the best (M=18.46.), followed by the mid-ability group (M=11.33), and then the low-ability group (M=8.02). The standard deviations for all three groups are closer together than with the dichotomous scoring method. The reliability estimate using this scoring procedure (α = .89) is higher than the results from the dichotomous scale (r = .83), indicating less measurement error.

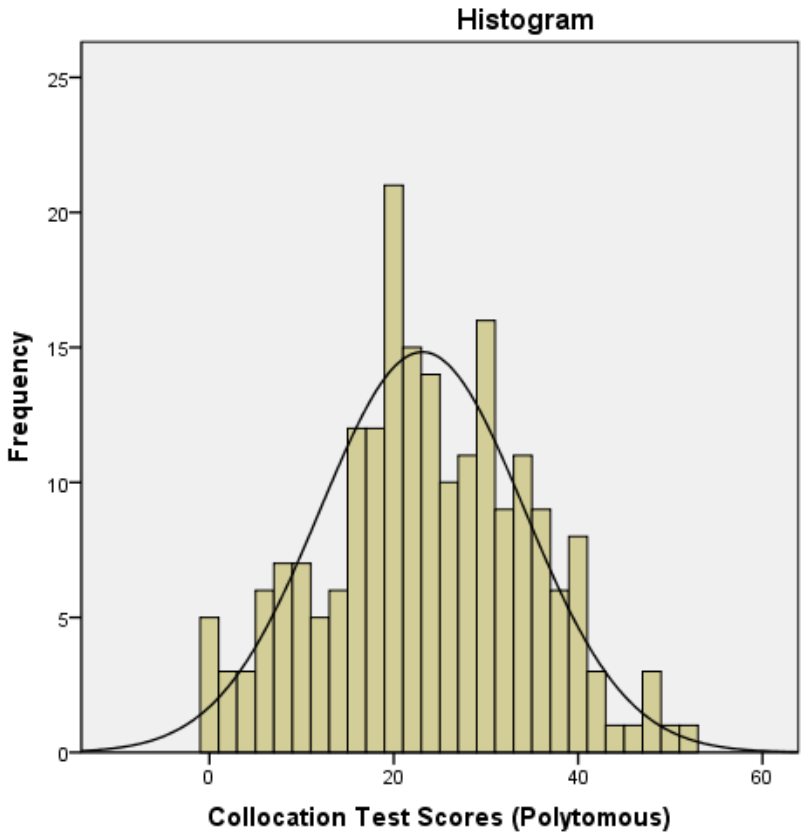

Figure 5.2. Histogram showing the distribution of scores on collocation test (k=35) for all groups (N=206)

Furthermore, scores using the polytomous scoring method are more normally distributed. The symmetrical distribution of scores presented in Figure 5.2 is more beneficial for making placement decisions provides backing for the assumption that scores had appropriate characteristics for norm-referenced decisions was supported for the polytomous scoring method.

In sum, the descriptive statistics indicate that the higher proficiency groups outperform the lower proficiency groups in a descending order. This is true for both scoring methods; however, the application of the polytomous scoring method produced a score distribution that is favorable to the dichotomous scoring method for making decisions for placing students into appropriate levels of English instruction or exempting them from instruction. The consistency of the measurement is superior for the polytomous scoring method, as indicated by the reliability estimates. Backing was also found for the evaluation inference using the polytomous scoring method only.

5.2 Rasch IRT model fit

The data from both scoring methods were evaluated for item fit using a Rasch model analysis with Bond and Fox Steps software (Bond & Fox, 2007). Rasch is considered equivalent to the one-parameter model (Hambelton & Swaminathan, 1985). Rasch estimates for both scoring methods are presented in Table 5.3.

Table 5.3. Standardized Rasch estimates for both scoring methods

INFIT ZSTD	Dichotomous	Polytomous
Mean logit	.1	-.1
SD logit	1.0	1.5
Highest ability	2.9	3.5
Lowest ability	-3.3	-3.4
Logit range	6.2	6.9
Rel. of estimate	.95	.97
Misfitting items	2	8

The Rasch analysis calculates item difficulties as a logit interval scale. The mean of item estimate in logits for both scoring methods is close to the standard mean of zero. There is a larger standard deviation with the polytomous data. Logit ranges for both methods are similar. All items are located within 2.9 and -3.3 logits for the dichotomous method with a range of 6.2 and between 3.5 and -4.3 logits for the polytomous method with a range of 6.9. The estimates for the items are highly reliable (.95 - .97).

Table 5.4 presents the items along with their infit and outfit statistics for both scoring methods. The data are presented as infit and outfit statistics. For this study, considering the small sample size, only the infit statistics were considered useful, because outfit statistics can be affected by a small number of outliers. Infit statistics are identified by the values of the mean square ranges (MNSQ) and the standardized ranges (ZSTD). Bond and Fox (2007) recommended that values for the mean square ranges should be between 0.4 and 1.12 for items that are judged; however, due to a smaller sample size, the upper value may reach 1.5. Items not within these parameters should be considered misfitting items. Standardized fit statistics (ZSTD) were used to identify misfitting items, because this statistic is not

dependent on sample size. Accordingly, standardized fit statistics for misfitting items are identified by t-values that are greater than +2 and less than -2 (Bond and Fox, 2007).

Following these criteria for identifying items that fit the model, there were two misfitting items in the dichotomously scored data and 8 misfitting items in the polytomously scored data. There is little variability in the infit statistics, indicating that the Rasch model was fitting well overall, considering the number of participants.

The two misfitting items in the dichotomous data are shown in Table 5.4. The two collocations were item 10, "have control," with the collocate form "has" on the test, and item 30, "make a decision," with the collocate form "made" on the test. Both of these collocations had the highest frequency of all target collocations. Whereas these collocations are extremely frequent, the likelihood that they are also frequent in domains other than academic English is high, indicating that they might not be representative of the items from an academic domain. This issue was taken into consideration by Coxhead (2000) when developing the academic word list by removing items that were not exclusive to the academic language domain from the initial academic frequency list. This study did not remove any collocations, since they were considered highly frequent in academic written English and would be representative of collocations in that domain.

The polytomous data analysis produced eight misfitting items: 1, 2, 5, 7, 20, 25, 28, and 31. These collocations were "have a chance" (have), "make a distinction" (made), "exert an influence" (exert), "make arrangements" (made), "have an understanding" (have), "have an effect" (have), "receive attention" (received), and "provide a service" (provided). The form of the collocate is indicated in parentheses.

Items with negative infit statistics in Table 5.4 are considered overfitting the model. That means that test-takers with low ability were not able to perform well on difficult items and that test-takers with high ability performed very well on difficult items. Items with positive infit statistics indicate that the data were underfitting the model. In some cases, test-takers with low ability were performing well on items with high difficulty, or test-takers with high ability were performing poorly on difficult items.

Table 5.4. Item response theory (IRT) fit statistics for both scoring methods

Dichotomous Scale						Polytomous Scale					
ITEM #	TOTAL SCORE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	ITEM #	TOTAL SCORE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
26	89	1	0.1	1.05	0.6	21	255	0.96	-0.5	0.97	-0.3
20	82	1.1	1.5	1.06	0.7	27	223	0.89	-1.5	0.83	-1.5
30	80	0.8	**-3.3**	0.72	-3.2	11	215	1.02	0.3	1.01	0
25	76	1.12	1.7	1.11	1.2	26	214	0.97	-0.3	0.92	-0.8
35	70	0.95	-0.7	0.99	-0.1	10	213	1.01	0.1	1	0.1
22	66	0.97	-0.4	0.96	-0.3	**31**	202	0.76	**-3.4**	0.72	-2.5
10	65	**1.23**	**2.9**	1.28	2.4	19	171	0.89	-1.4	0.85	-1.7
9	59	1	0	0.93	-0.5	12	169	1.02	0.2	1	0.1
7	46	0.93	-0.8	1.08	0.5	**20**	162	0.82	**-2.2**	0.79	-2.4
18	46	0.88	-1.3	0.8	-1.3	23	159	0.97	-0.4	0.89	-0.5
6	43	1.13	1.3	1.4	2.1	4	158	1.06	0.8	1.07	0.4
31	36	1.04	0.4	0.99	0	22	157	0.9	-0.9	0.81	-1.5
24	35	0.84	-1.5	0.76	-1.2	32	152	0.95	-0.6	0.94	-0.6
15	34	1.04	0.4	1.13	0.7	9	151	1	0.1	0.99	0
29	33	0.98	-0.1	0.72	-1.4	**7**	146	1.19	**2.2**	1.41	3.1
17	33	0.89	-0.9	0.75	-1.2	**5**	146	1.23	**2.6**	1.24	2.2
3	31	1.02	0.2	0.92	-0.3	18	135	0.9	-1.2	0.84	-1.4
4	30	1.15	1.2	1.41	1.6	**25**	129	0.82	**-2.3**	0.75	-2.1
19	30	1.01	0.1	1.04	0.3	34	127	1	0	1	0.1
33	29	0.99	0	0.89	-0.4	14	124	0.95	-0.5	0.92	-0.8
32	26	0.91	-0.6	0.82	-0.6	8	123	1.07	0.8	1.32	1.7
1	17	1.25	1.3	1.93	2.2	17	108	0.9	-1.2	0.86	-1.4
2	15	0.96	-0.1	1	0.1	6	106	1.08	1.1	1.09	0.9
8	13	0.87	-0.5	0.72	-0.6	16	102	1.03	0.4	1.24	1.2
11	11	1.05	0.3	1.72	1.4	33	98	0.89	-1.1	0.84	-0.9
13	8	0.99	0.1	0.96	0.1	30	96	1.04	0.4	1.08	0.5
16	8	0.98	0	0.56	-0.7	**28**	96	0.82	**-2.3**	0.76	-2.4
34	8	0.97	0	0.65	-0.5	**2**	92	1.38	**3.4**	1.87	4.7
14	7	1.05	0.3	1.05	0.3	29	83	0.98	-0.2	0.91	-0.9
12	6	0.85	-0.3	0.36	-1.1	**1**	61	**1.55**	**3.5**	2.29	5.2
27	5	1.02	0.2	0.48	-0.7	15	60	0.97	-0.2	1.07	0.3
23	5	0.94	0	0.43	-0.8	35	56	1.09	0.7	1.03	0.2
5	4	0.99	0.1	0.63	-0.3	24	50	1.01	0.1	1.04	0.2
21	3	1.05	0.3	1.95	1.2	3	36	1.03	0.2	3.75	3.5
28	2	1	0.2	0.44	-0.5	13	29	1.01	0.1	2.26	2.7

An item characteristic curve (ICC) is a model of an item indicating the parameters that are being measured. Since this is a one-parameter IRT model, only the b-parameter, *difficulty*, is shown. The a-parameter, *discrimination*, is held constant, and the c-parameter, *guessing*, is ignored. Each curve represents a standardized mean of zero on the horizontal axis and a standardized probability of obtaining a correct answer as 0.5 on the vertical axis. As the difficulty increases, the probability of getting a correct answer also increases.

Figures 5.6a and 5.6b show the item characteristic cures for the two misfitting items in the dichotomous data. The model curve is shown in red and the empirical curve is presented in blue with black points. The black points represent groups of test-takers with similar abilities. Test-takers are grouped by the software and vary in number in each group. A perfect fit would show the empirical line with groups following the red model line for each item.

Have control (has)

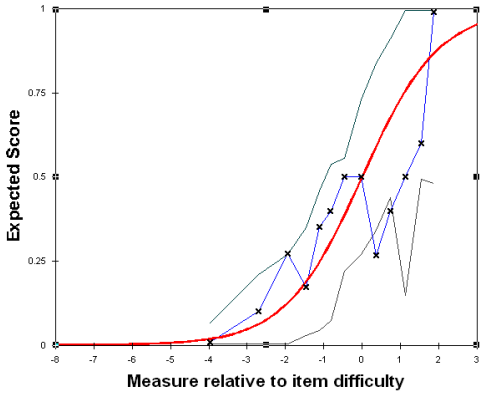

Figure 5.3a. ICC for item 30 (b=2.9)

Make a decision (made)

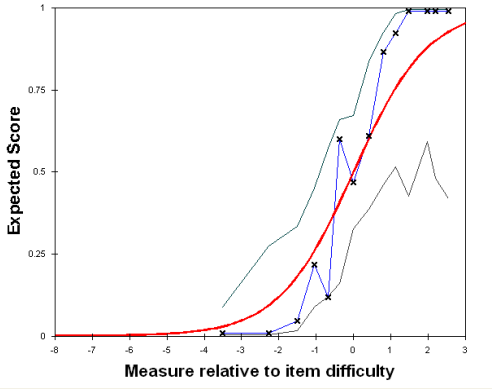

Figure 5.3b. ICC for item 10 (b=-3.3)

Figure 5.3a and 5.3b show the item characteristic curves for the first of two misfitting items. Item 30 (have control) has an underfitting t-score (INFIT ZSTD) of 2.9. Item 10 (make

a decision) has an overfitting t-score statistic of -3.3. An overfit statistic indicates that the item performs better than the model would predict. The data in this case are too predictable. Test-takers with lower ability would not provide a correct response for this item and test-takers with higher ability would provide a correct response. Item 30 appears to have better discrimination than expected. This is not surprising since "Make a decision" is such a frequent collocation. Performance on Item 10, on the other hand, is not as expected between 0 and 2 on the horizontal ability scale. Test-takers with higher ability were not performing as well as predicted apart from the group with the highest ability. Whereas these data represent the scores using the dichotomous method, one interpretation could be that the lower ability test-takers were able to produce the target collocation slightly better than expected. Guessing might have played a role in these responses since the target response was a frequent delexicalized verb "has" The highest ability test-takers were able to produce the expected response. These test-takers might be familiar with the collocation and were able to recognize it in context. The groups in the middle of the empirical line, representing test-takers with moderate ability, might be acquiring a sense of academic language and were experimenting with their response to produce a collocate that appeared to be more academic than the target collocate "have." A number of responses for this item (e.g., achieve, gains, exert, enforces, implement, impose, lose) received partial credit with the polytomous method but not with the dichotomous method. The collocation "lose control" is a common restricted verb-noun collocation that was not awarded any credit with the dichotomous method, because it was not the correct response. The response "lose" would not have been awarded full credit with the polytomous method either, because it was not appropriate for the context, but it did receive partial credit as evidence of partial knowledge.

Item 10

For the most part the school _____ little control over these types of evaluation so we will consider the approaches in outline, concentrating on the issues raised. [key: has]

Items 10 and 30 would not have been identified using only CTT item-total statistics. Item-total statistics for all items are presented in Appendix F. The reliability estimate would have remained the same (α= .83) if item 10 was removed and would have decreased (α = .82) if item 30 was removed.

Figures 5.8a through 5.8h show the item characteristic curves for each of the eight misfitting items using the polytomous scoring method. Figures 5.4a–5.4d present ICCs for the overfitting items. Figures 5.4e–5.4h show ICCs for the underfitting items. It is clear to see from the ICCs which items are overfitting (items 20, 25, 28, and 31) and which are underfitting (items 1, 2, 5, and 7). If overfitting items are considered acceptable, there are only 4 items out of 35 that may cause trouble for a test-taker and should be reviewed. ICCs are shown for all misfitting items, but because overfitting items are very predictable and fit the model better than expected only the underfitting items are addressed.

have an understanding (have)

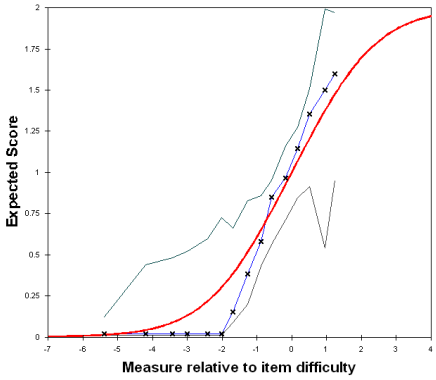

Figure 5.4a. ICC for item 20 (-2.2)

provide a service (provided)

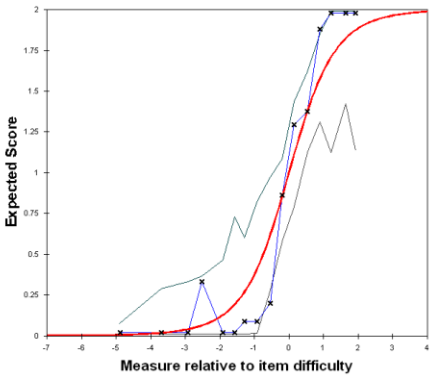

Figure 5.4b. ICC for item 31(-3.4)

have an effect (have)

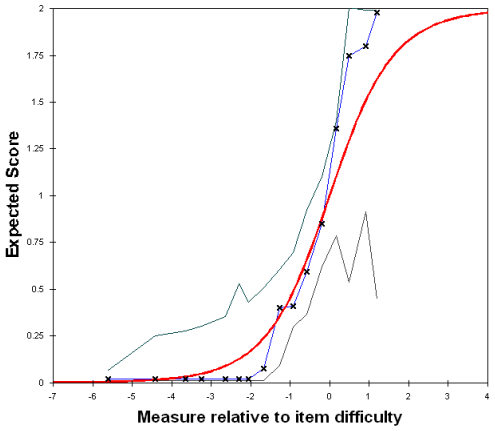

Figure 5.4c. ICC for item 25(-2.3)

receive attention (received)

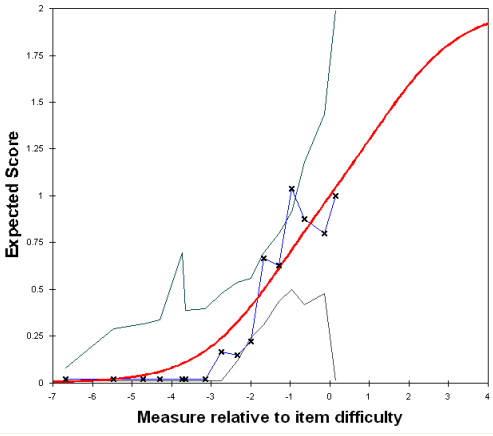

Figure 5.4d. ICC for item 28 (-2.3)

have a chance (have)

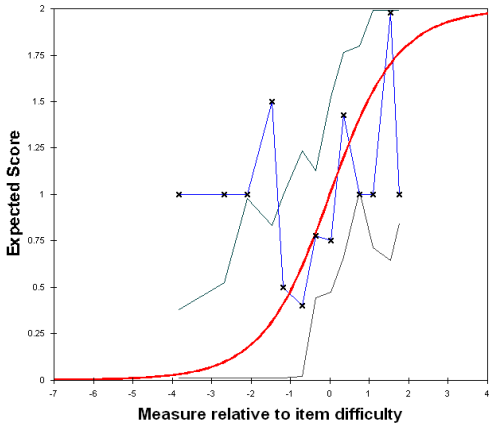

Figure 5.4e. ICC for item 1 (b=3.5)

Item 1

A curriculum-based report is likely to _____ a greater chance of success because it shows benefits to pupils, the teachers, and the school librarian i.e., the whole school. [key: have]

Correct (D)/ Full Credit (P): have, has
No Credit (D)/ Partial Credit (P): e.g., take, offer, increase, find, give
No Credit (D and P): e.g., guarantee, illustrate, lead, introduce, meet

Item 1 appears to be misfitting because of the high percentage of low performing test-takers who provided a correct response to this item. Sixty-nine different responses were provided by test-takers. Only two were target collocations and 11 response types were awarded partial credit. All other responses were not given any credit. The high percentage of correct responses by low performing test-takers could be attributed to guessing since the target collocate is a highly frequent delexicalized verb, "have."

make a distinction (made)

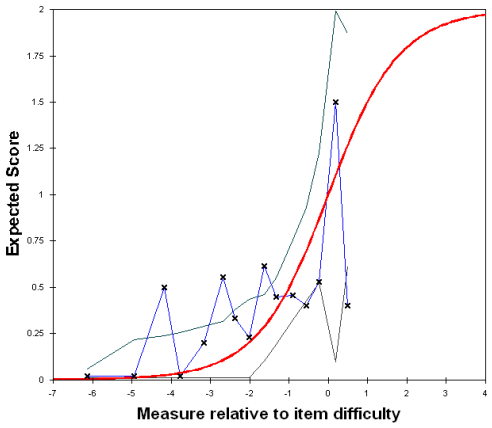

Figure 5.4f. ICC for item 2 (b=3.4)

Item 2

A distinction can be _____ between planned and unplanned decentralization. [key; made]

Correct (D)/ Full Credit (P): made, make, makes, maken, making
No Credit (D)/ Partial Credit (P): e.g., consider, created, provided, found
No Credit (D and P): e.g., known, identified, selected, chosen

According to Figure 5.4f, the empirical data show that some low performing test-takers were providing correct responses for Item 2, whereas the top performing test-takers were not responding correctly. One-hundred fifteen different response types were recorded for this item. Correct or full credit was awarded to five different forms. Partial credit was

given to 17 response types. A combination of the two strategies mentioned can be seen already. The low performing test-takers might be guessing correctly, because the verb "make" is a highly frequent verb. The high performing test-takers may be experimenting with words that are perceived to be more academic, as indicated by the responses that received partial or no credit.

exert an influence (exert)

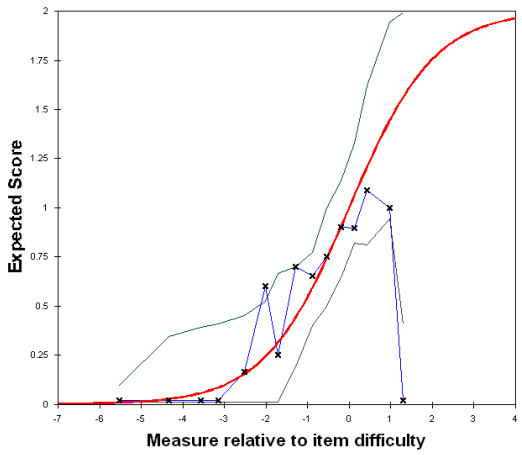

Figure 5.4g. ICC for item 5 (b=2.6)

Item 5

Although it might be very easy to steal from a friend or colleague and not get caught most people would feel this to be "wrong", yet such feelings may not _____ such a strong influence over decisions as to whether to steal from a larger and less personal victim. [key: exert]

Correct (D)/ Full Credit (P): exert
No Credit (D)/ Partial Credit (P): e.g., feel, have, establish, create
No Credit (D and P): e.g., trigger, cost, decrease, produce

There were seventy-three total response types for item 5. Thirty response types were awarded partial credit. The misfitting data for this item are seen at the upper end of the scale with the higher preforming test-takers. The target collocate for this collocation is unique, as it

is not a common word but instead a verb that might not be known by most test-takers. This being the case, it is uncertain if the highest performing test-takers were not familiar with this verb or if they were experimenting with verbs that they perceived to be academic. Another interpretation is that these test-takers noticed the high number of target collocations that were correct with "have" and "make" and believed that most of the items did contain delexicalized verbs rather than a single academic verb such as "exert" rather than produce a verb that the test-taker perceived as more academic. Test-takers thus may have been guessing by producing a highly frequent verb instead of a less frequent verb even if they were familiar with the verb "exert."

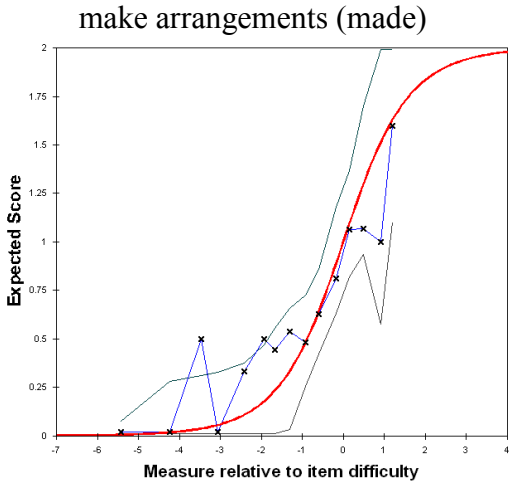

Figure 5.4h. ICC for item 7 (b=2.2)

Item 7

Claudius _____ two other arrangements which seem at first sight to be highly anomalous. [key: made]

Correct (D)/ Full Credit (P): made, make, makes
No Credit (D)/ Partial Credit (P): e.g., had, has, includes, create, consider
No Credit (D and P): e.g., chose, finish, opposed, promote

There were ninety-five total response types for item 7, 15 of which received partial credit with the polytomous scoring method. This item had the largest misfit of the model, with the lower performing test-takers between -4 and -3 on the difficulty scale. Once again, many of the lower performing test-takers were potentially guessing correctly with a target response which could be attributed to the high frequency of the collocate for this item. There were a considerably larger number of unique responses for this item; however, apart from the large spike at the lower end of the scale and the slight deviation at the top, the empirical data are fairly close to the model predicted by the Rash analysis.

After examining the item characteristic curves for the underfitting items for both scoring methods, it appears as though test-takers may have been guessing by producing high-frequency verbs such as “have” and “make.” Since these verbs were target collocates for a number of collocations on the test, credit was awarded for these responses, particularly with the polytomous scoring method. It is also possible that test-takers may have taken a different approach and experimented with verbs which were considered more “academic” than the high-frequency verbs. Verbs such as “have” and “make” might not sound formal enough for such a test. As a result, the test-taker was given partial or no credit, because the response was not acceptable in written academic English or not frequent enough to have been selected as the target collocate.

More items using the dichotomous scoring method fit the Rasch model better than items using the polytomous scoring method. Nonetheless, the percentage of misfitting items for both scoring methods was relatively small. Misfitting items need further investigation but were not necessarily performing so poorly as to remove them from the test.

5.3 Reliability

Findings from the second research question are needed to support the first assumption for the generalization inference that a sufficient number of items were included on the test to provide reliable estimates of test-takers' performance. This question asks whether a test constructed using high-frequency collocations in written academic English can (a) reliably distinguish among test-takers and (b) produce scores with acceptable reliability. The answer to this question was found with results from a Rasch IRT analysis and the descriptive statistics presented above for both scoring methods. The results that would provide the evidence necessary with an IRT analysis are larger reliability values for person separation. Person separation values indicate the degree to which the scores on the test can distinguish among test-takers. The person reliability associated with the separation values is interpreted on a scale from 0 to 1. Larger reliability values indicate more consistent measurement.

A second value that indicates consistency in measurement is the reliability value for the test itself as measured by Cronbach's alpha. Considering the sample size and number of items on the test, a reliability estimate equal to or greater than .80 would be needed to be acceptable for the uses of the test (Bachman, 2004; Carr, 2011).

The values for person separation, separation reliability, and Chronbach's alpha reliability are shown in Table 5.5. Results show that the polytomous scoring method distinguishes more reliably among test-takers (.69) than the dichotomous method (.87).

Table 5.5. Rasch person separation and reliability estimates for both scoring methods

Scoring method	Person separation	Person reliability	Cronbach's alpha
Dichotomous	1.48	.69	.83
Polytomous	2.59	.87	.89

A similar result was found with Cronbach's alpha. Reliability estimates reveal that both scoring methods produced reliability estimates above .80. The scores using the polytomous scale were found to have a higher reliability estimate ($\alpha = .89$) than the scores using the dichotomous scoring method ($\alpha = .83$). Approaching an estimate of .90, the reliability estimate for the polytomous data produces more consistency in the scores, indicating a reduction in measurement error.

Both the IRT analysis and the CTT reliability estimates provide backing for the assumption that a sufficient number of items are included on the test to provide reliable estimates of test-takers' performance. This assumption supports the warrant that expected scores are comparable across the items on the test. Backing for this warrant allows us to move from the generalization inference to the explanation inference in the validity argument.

5.4 Item facility and rank order

The next research question was developed to seek backing for the third assumption supporting the explanation inference. This assumption states that the more frequent a collocation is, the easier the corresponding item will be for test-takers. This relationship is generally assumed to be true, based on the theory of frequency discussed in chapter 3. The theory is that test-takers should perform better on test items that are based on frequent collocations. The interpretation of this notion could be related to the item difficulty. Theoretically, items that contain frequent collocations are easier for the test-takers than items with less frequent collocations. Backing for this assumption will support the underlying warrant that expected scores are attributed to a construct of collocational ability in academic writing, which includes knowledge of whole collocations.

This research question investigates whether item facility, the difficulty of an item, is related to the frequency of occurrence of a collocation in the target domain. In other words, are more frequent collocations easier for test-takers? In order to answer RQ 3, correlations were calculated between the item means for each scoring method and the rank order of the collocations, from most frequent to least frequent. Table 5.6 shows the rank order of the collocations by descending frequency, as identified in the corpus of general written academic English. The list shows collocations consisting of the verb, which is the collocate, and the noun, which is the node, followed by the form of the collocate used on the test in parentheses. The list is sorted by raw frequency of the collocations in the corpus of written academic English.

Table 5.6. Rank order of collocations by frequency in the corpus of written academic English

Rank order	Collocation	Raw frequency	Rank order	Collocation	Raw frequency
35	Have an effect (have)	2484	17	Give notice (given)	276
34	Make a decision (made)	1175	16	Make reference (made)	255
33	Play a part (play)	801	15	obtain information (obtained)	252
32	Provide service (provided)	682	14	Make effort (made)	241
31	Have control (has)	629	13	Have understanding (have)	239
30	Make an attempt (made)	551	12	Make contract (made)	234
29	Pay attention (paid)	503	11	Give attention (given)	229
28	Use technique (using)	481	10	Have chance (have)	217
27	Use term (used)	473	9	Make a choice (made)	212
26	Have access (have)	464	8	Reach a conclusion (reached)	208
25	Take step (taken)	452	7	Make difference (make)	199
24	Take care (taken)	413	6	Obtain results (obtained)	164
23	Take form (take)	412	5	Give consideration (given)	164
22	Take action (take)	404	4	Make statement (made)	163
21	Make distinction (made)	351	3	Receive attention (received)	151
20	Make contribution (make)	346	2	Make arrangements (made)	141
19	Treat patient (treated)	329	1	Exert an influence (exert)	89
18	Collect data (collected)	317			

The raw frequency count in the corpus for each collocation is also shown in Table 5.6. The range of raw frequency is from 89 to 2484 hits in 16 million words. Normalized values range from 5.5 to 154.5 per million words. These frequency counts were used to develop the test based on the notion of item development using high-frequency test items. Since collocations are not as frequent as individual words, the raw frequency counts may appear to be low, although the sample of collocations comes from the most frequent word pairs identified by the software analysis. Many collocations with lower frequency counts were not selected for the test because they were so infrequent relative to the other collocations identified in the target domain.

5.4.1 Correlation between rank order and item facility with dichotomous scoring

A correlation was calculated between the rank order of the collocations and the item means of the scores using the dichotomous scoring method. Because of the rank-ordered data involved, a Spearman's Rho correlation was carried out on rank order of the collocations from the corpus and item facility of the same collocations on the test. The results indicated almost no relationship between the relative frequency of a collocation, as a word pair, and the item facility of the corresponding items on the test ($r_s = .02$) using the dichotomous scoring method. This correlation between rank order of rank order and item facility is shown in the scatterplot in Figure 5.5

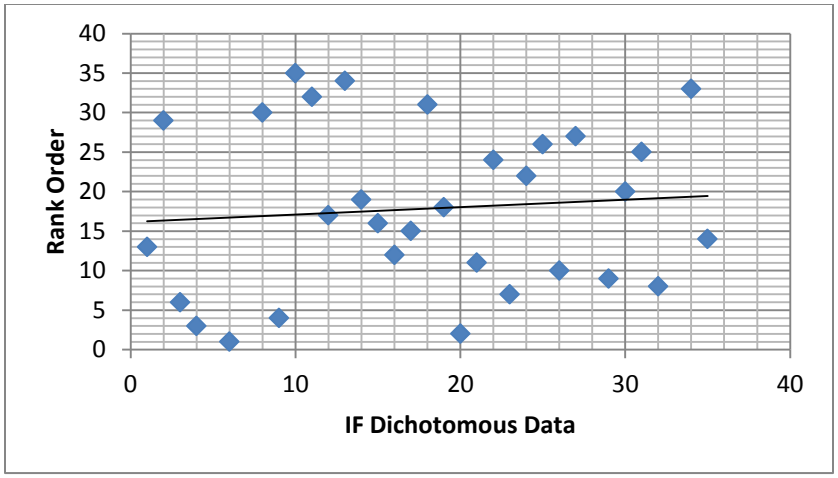

Figure 5.5. Scatterplot showing the relationship between rank order for collocation frequency and item facility using the dichotomous scoring method (r_s = .02)

The correlation between rank order of items based on frequency and item facility for the test scores have shown almost no relationship. These results may indicate that frequency does not play a large role in item facility relative to the rank order in the corpus. This interpretation is based on rank order of collocations from a sample which are all considered highly frequent collocations. The correlational coefficient was thus an indication of differences in item facility on a very small scale, which stopped at 89 occurrences. Collocations could have been sampled with low frequencies down to about five occurrences in the entire 16-million-word corpus. The results might have been different by including collocations with a larger range of item frequency; however, the range in frequency for the sampled collocations was limited to the group of highly frequent collocations, following the specification of the test calling for items to be developed using high-frequency collocations.

5.4.2 Correlation between item frequency and item facility with polytomous scoring

In this section, the relationship between rank order of collocations by frequency and item facility with the polytomous scoring method is presented. Because of the ordinal data using rank order of item frequency, the Spearman Rho correlation coefficient was used to calculate the relationship. The correlation coefficient indicated a moderate yet significant relationship between frequency and facility (r_s = .40). A scatterplot showing the relationship between rank order for collocation frequency and item facility using the polytomous scoring method is presented in Figure 5.6.

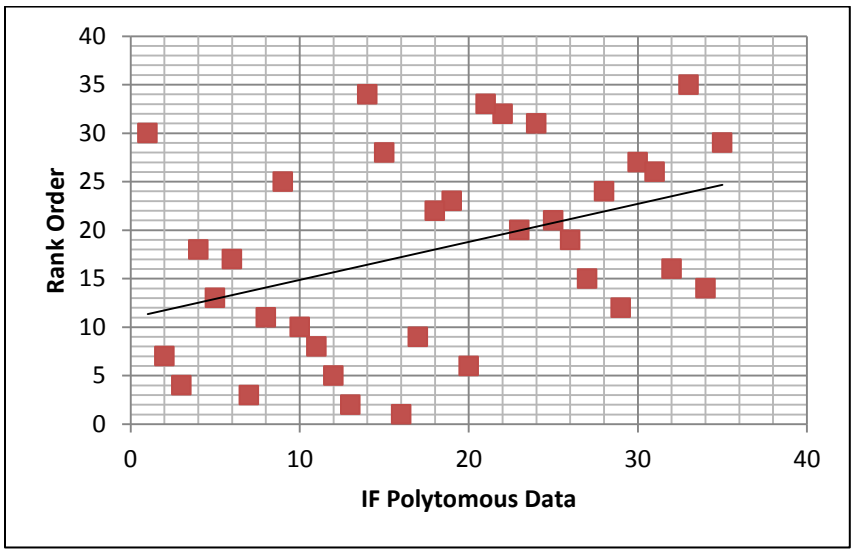

Figure 5.6. Scatterplot showing the relationship between rank order for collocation frequency and item facility using the polytomous scoring method (r_s = .40)

The correlation between rank order and item means was stronger using the polytomous scoring method than the dichotomous method. This scoring scale appeared to be more sensitive to the rank order of sampled collocations from the corpus. Even with the smaller range in frequency, a relationship begins to be evident between item facility and item frequency with the polytomous scale. Furthermore, Williams' test was used to test the null

hypothesis for correlations with dependent correlations (Steiger, 1980). Results indicated that correlation coefficients for dichotomous and polytomous scoring methods were significantly different.

5.5 Nomological network

Research question four seeks to support the assumption that test-takers' performance on the collocation test correlate positively with performance on concurrent measures of reading and vocabulary size as expected. Positive relationships were expected because of the network of constructs, which form connections in the nomological network as discussed in chapter 2. This assumption, in turn, supports the warrant that expected scores are attributed to a construct of collocational ability in academic writing, referring to knowledge of a whole collocation. This evidence provides backing for the explanation inference.

The theoretical relationships among reading, vocabulary, and collocations established earlier laid the foundation for the expected correlations among these measures. Because collocations are combinations of vocabulary, a positive relationship with a measure of vocabulary size would be expected. Production of collocations, however, involves the ability to combine words in context in a different way than with individual lexical items, which are not restricted in their potential combinations. The relationship would thus be expected to be moderate. The expected relationship with reading and collocational ability would also be moderate for a couple reasons. First, collocations appear in many forms in written language and either add to the difficulty of a reading text or makes the text easier. A better reader would be more familiar with a variety of word combinations. On the other hand, reading for meaning is different than producing language. A characteristic of the verb-noun collocations

on this test is that one of the constituents may lose its primary meaning, as the meaning of the collocations are not always noticed since it is the node that is the meaning bearing unit of the collocation and the verb is often overlooked. On a test of collocational ability, the ability to produce a target or appropriate collocate rests on the knowledge of the word pair and the ability to produce the combination in the correct context. Based on the theory presented in chapter 2, one would anticipate a correlation between reading and a measure of vocabulary size to be fairly high.

5.5.1 Descriptive statistics for reading test

Descriptive statistics for the reading test for all three groups (k=35) are presented in Table 5.7. The scores for the reading test are non-normally distributed, with a moderate reliability estimate (α = .78). The maximum possible score on the reading test was 20. The mean scores follow the expected pattern, with the high-ability group performing best (M= 18.46), followed by the mid-ability group (M=15.22), and finally the low-ability group (M=13.25).

Table 5.7. Descriptive statistics for the reading test for all levels (k=35)

Level	N	Min	Max	Mean	Std. deviation	Cronbach's alpha
High-ability	35	15	20	18.46	1.48	
Mid-ability	109	1	20	15.22	3.12	.78
Low-ability	62	4	20	13.25	3.99	
TOTAL	206	1	20	15.18	3.63	

A one-way ANOVA indicated that the three groups were significantly different, $F(2,203) = 29.26$, $p < .000$, which was supported by a Tukey HSD Post-Hoc Test. A

histogram showing the distribution of scores on the reading size test by all groups is presented in Figure 5.7.

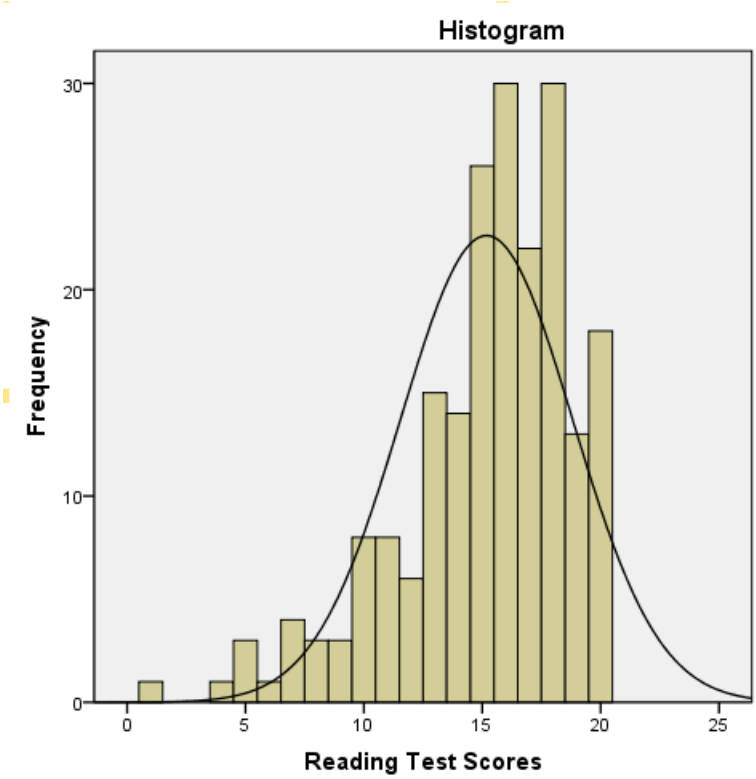

Figure 5.7. Histogram showing distribution of scores on reading size test by all groups (k=20)

5.5.2 Correlations with reading test

A significant relationship was found for the total scores on the Collocational Ability Test using both the dichotomous and polytomous scoring methods and the general reading test. The relationship between scores on the collocation test and scores on the reading test for both scoring methods are shown in Figures 5.8 and 5.9 respectively. These relationships were positive but not very strong. Spearman's Rho formula, which does not assume normality, was used to calculate the correlation coefficient between the scores on the Collocational Ability Test using the dichotomous scoring method and the scores on the reading test. This moderate

relationship (r_s = .70) was higher than the relationship between the polytomous Collocational Ability Test scores (r_s = .62). Both correlations were corrected for attenuation. The correlational evidence provided by these comparisons indicates that the abilities measured are different. Whereas the correlation with the dichotomous scale is higher than the polytomous, there may be more in common with reading and performance on the test as scored with the dichotomous scoring method. This relationship is likely, because the more proficient readers were able to identify a correct collocate, whereas other test-takers were given no credit for partially correct responses resulting from possible guessing. Word combinations and collocations can influence the difficulty of a reading text. The attenuated correlations show that there is a moderate relationship between reading ability and productive collocational ability.

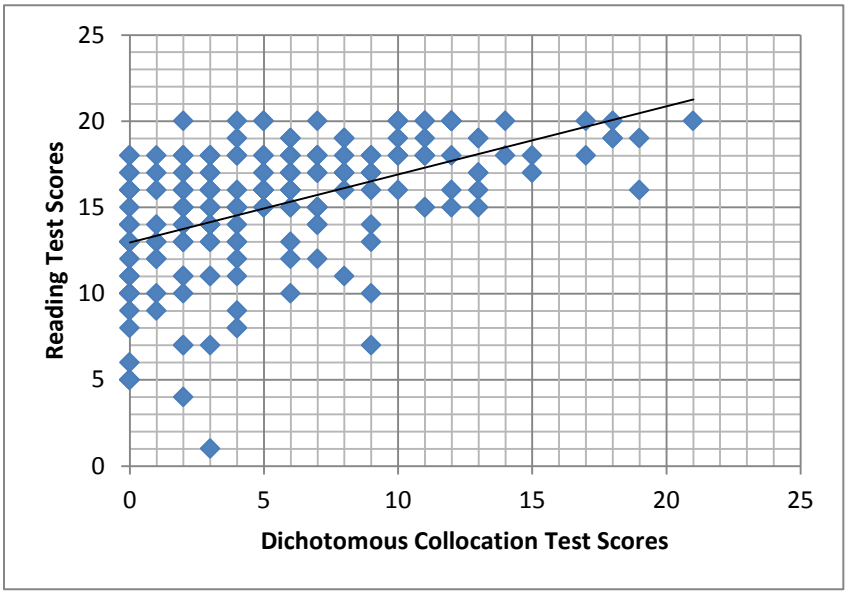

Figure 5.8. Scatterplot showing relationship between dichotomous collocation test scores and reading test scores (r_s = .70)

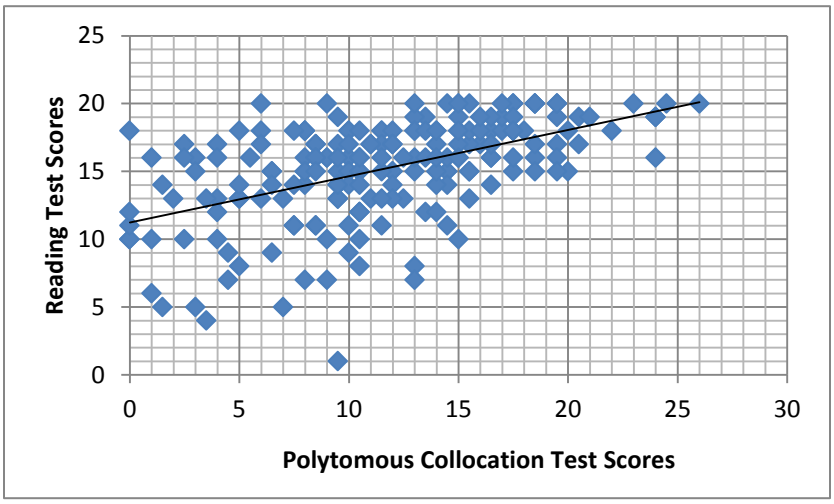

Figure 5.9. Scatterplot showing relationship between polytomous collocation test scores and reading test scores (r_s = .62)

5.5.3 Descriptive statistics for productive vocabulary size test

Descriptive statistics for the scores from the vocabulary size test (k=30) in each of the three groups are shown in Table 5.8. The mean scores confirm the predicted patterns with the high-ability group obtaining the highest mean score (M=23.92) and decreasing as expected with the other groups. The mean score for the mid-ability group was fairly high (M=18.24), and the mean score for the three groups of test-takers in the low-ability group was lower. The skewness and kurtosis statistics are not extreme, but a Shapiro-Wilk W "goodness-of-fit" test showed the scores on the vocabulary test to be non-normally distributed. The non-normal distribution is supported visually by the histogram in Figure 5.10.

Table 5.8. Descriptive statistics for all three groups on productive vocabulary size test (k=30)

Level	N	Range	Min	Max	Mean	Std. deviation	Cronbach's Alpha
GR	35	24	5	29	23.92	4.17	
UG	109	25	2	27	18.24	4.73	.87
IE	62	21	3	24	13.21	4.84	
Total	206	27	2	29	17.78	5.86	

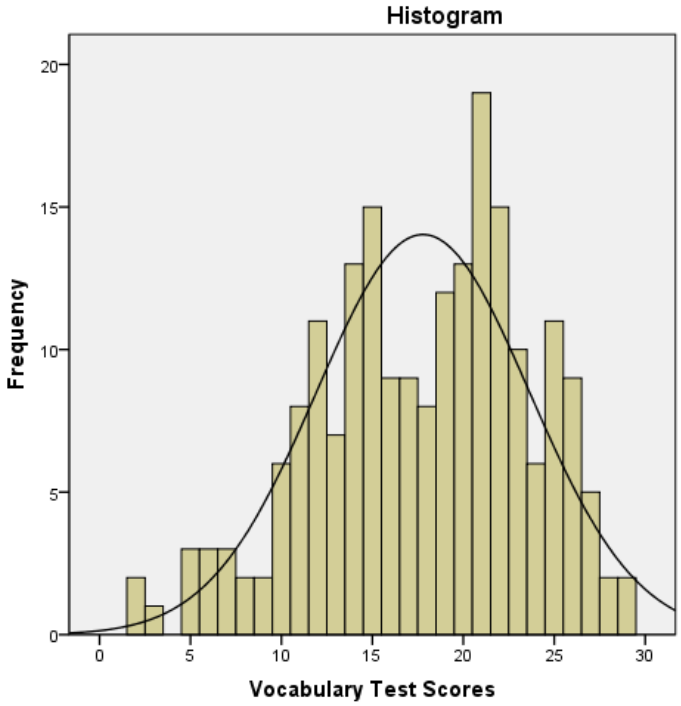

Figure 5.10. Histogram showing distribution of scores on vocabulary size test by all groups (k=30)

5.5.4 Correlations with vocabulary size test

Assuming that measures of vocabulary size and depth develop simultaneously as proficiency level increases, as suggested by Akbarian (2010) and Vermeer (2001), test-takers' performance on the collocation test would be expected to correlate positively with performance on measures of overall productive vocabulary size. One would predict that

frequent lexical items in a specific discourse domain are encountered and possibly produced by language learners as well as target language users so often that the words in the collocation are commonly co-selected when language is produced in a familiar target language context. Language learners with greater vocabulary ability thus would be more likely to produce an acceptable collocation as well as appropriate individual lexical items in free production; however, the correlation coefficient that is too strong would suggest that the collocational ability test and the vocabulary test are measuring the same construct and do not separate vocabulary size and depth.

In addition, the relationship between performance on an academic reading test and performance on a vocabulary depth test should be similar to performance on a vocabulary size test (Qian, 2002). For both scoring methods, we would expect a high correlation with a vocabulary size test if it is true that collocational ability develops along with vocabulary size. The scatterplot in Figure 5.11 shows a positive relationship between scores on the vocabulary test (size) and scores on the collocation test (depth) using the dichotomous scoring scale and in Figure 5.12 using the polytomous scoring method.

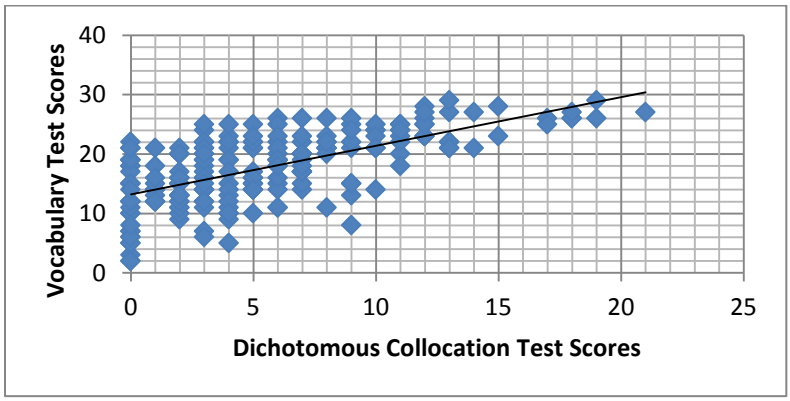

Figure 5.11. Scatterplot showing the relationship between Collocational Ability Test scores using dichotomous scoring and vocabulary size test total scores (r_s = .74)

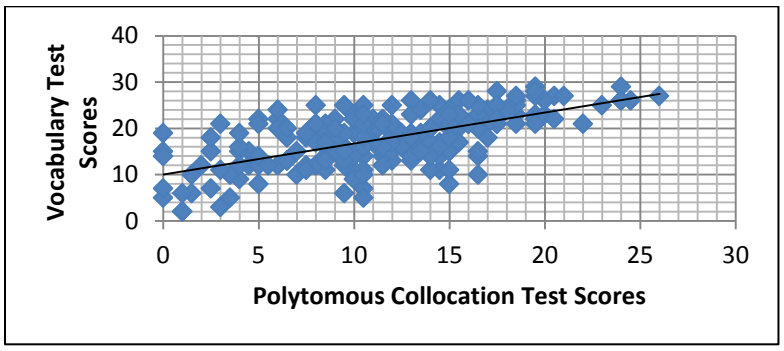

Figure 5.12. Scatterplot showing the relationship between Collocational Ability Test scores using polytomous scoring and vocabulary size test total scores (r_s = .72)

Attenuated correlations were calculated to compare the theoretical relationships between the constructs. Results confirm the expectations showing a moderate positive relationship for the collocation test scores and the vocabulary size test scores for the dichotomous data (r_s = 0.74) and for the polytomous scores (r_s = .72); therefore, collocational ability is positively associated with vocabulary size. In fact, the test results indicate that the relationships are statistically significant.

Relationships among theoretically related constructs in the nomological net were investigated using correlation coefficients corrected for attenuation. A Spearman's Rho correlation coefficient for reading and vocabulary size was found to be as predicted (r_s = .80). Higher correlations were found using the dichotomous data with reading (r_s = .74) and vocabulary size (r_s = .72) than the polytomous data with reading (r_s = .70) and vocabulary size (r_s = .62). Allowing for partial credit may change the initial construct, allowing for other knowledge and/or strategies to affect performance on the test. The predicted relationships among the theoretical constructs were found, however, as presented in chapter 2. The highest correlation was between reading and vocabulary size, followed by vocabulary size and

vocabulary depth. Finally, the relationship between reading and vocabulary depth was found to be the lowest, as predicted. The weaker relationship between reading and vocabulary depth might strengthen the claim that the Collocational Ability Test is, indeed, measuring collocational ability and is less affected by reading ability.

5.6 Strategies and perceptions

This section presents an empirical analysis and discussion of evidence supporting the fifth assumption in the explanation inference. This assumption states that test-takers use metacognitive and cognitive strategies that are related to collocation use in academic language while taking the test. Test-takers' perception of academic collocations and the language on the test were explored as part of the investigation into the use of metacognitive and cognitive strategies. The findings from these analyses can be used to answer RQ 5, showing the correspondence between students' perception of academic collocations and academic language and their performance on the Collocational Ability Test.

The analysis began by exploring potential strategy use by a sample of participants. Data was collected using screen capturing software. An analysis of the response pattern by the test-takers on the test indicated evidence of strategy use as test-takers changed their responses while they took the test. Simply producing a single response for each item without reflection was considered minimal use of metacognitive or cognitive strategies. Further evidence of strategy use was uncovered in the responses on the Test Reflection Survey. Responses were analyzed for evidence of familiarity with academic language on the test and in university texts. Responses by the test-takers who are familiar with university language should be able to indicate characteristics found in such texts and how the language on the test

was similar or different from academic English with which they were familiar. Finally, comments during post-test interviews were used to confirm or explain trends that were uncovered in the preliminary analyses. Backing for this assumption will supports the warrant underlying the explanation inference by showing that expected scores are attributed to a construct of collocational ability in academic writing.

5.6.1 Screen capturing data analysis

The video files from a sample of test-takers (n = 8) were analyzed and observations of the test-takers were recorded. The participants, their group membership, and scores on the Collocational Ability Test using both scoring methods are shown in Table 5.9.

Table 5.9. Participants involved in screen capturing data collection

ID#	Group	CAT score (Dich)	CAT score (Poly)	Response change	Mean
465	High	19	24	2	4.75
1112	High	13	19.5	4	
1107	High	13	20.5	6	
1108	High	10	16	7	
1071	Mid	4	10	3	2.5
1066	Mid	3	9	2	
609	Low	2	8	4	2.5
931	Low	0	1	1	

An analysis of the approach to the test by each test-taker revealed a few differences among the three levels. These changes are used as evidence of metacognitive strategies used while taking the test. The biggest difference is the number of times a response was changed by a test-taker. Average response changes by group are shown in Figure 5.13.

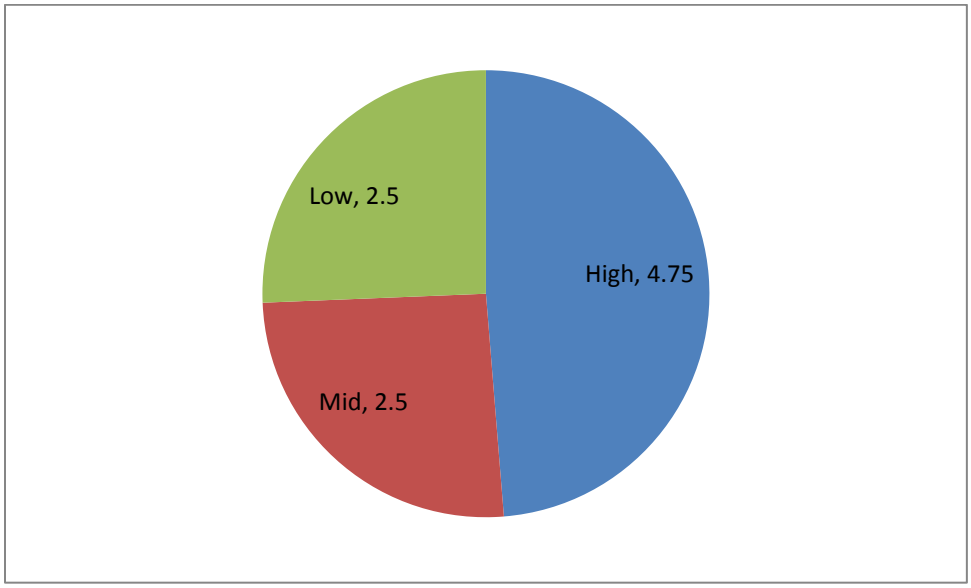

Figure 5.13. Average response change in percent by group (n = 8)

The high-ability group appeared to be more cognitively engaged in the process of identifying the context changing answer more frequently (M=4.75) than the mid-ability group (M=2.5) or the low-ability group (M=2.5). The alterations to the responses were more than just changing a base form verb to another form in order to match the appropriate grammar. The high-ability group made changes using lexical verbs such as the following: found > relevant > collected.

Other indications of the reading process were observed. Test-taker ID1112 was observed highlighting part of the sentence with the curser during reading. No test-taker, however, was observed using the cursor to identify a potential node for the missing collocate (verb). A second indication of the reading process is the sequence of responding to the items. The high-ability test-takers would start with item on and continue through to the bottom with a few exceptions. The low-ability test-takers would jump around from item to item. This also

resulted in a greater number of items with no response. It was not clear whether an item was skipped due to reading difficulty or inability to identify an appropriate collocate response.

A second difference is that that the high- and mid-ability groups would pause and scroll to the top of the page to read the instructions. This may or may not have assisted in helping the test-takers to focus on academic language.

Another indication of a test-taker's metacognitive process is revealed in the temporary response by test-taker ID1032 in the high-ability group. This test-taker indicated the base form of the verb she wanted to use but could not think of the appropriate form, so she marked it with a (pp) indicating a past participle form of the verb was needed. This response sequence for this item was: drew > draw(pp) > drown. She returned to the item later to changed her response to the correct form and came as close as she could.

Overall, the test-takers in the high-ability group were observed using more strategies than the other two groups. Test-takers in this group would systematically respond to the items in order. They would pause at times, scroll to the top of the test, and read the instructions and also change their answers as they engaged in responding to an item. Test-takers in the mid-ability group were also observed reading the instructions but did not change their responses as often. No test-taker in the low-ability group was observed reading the instructions. Test-takers in this group would jump around during the test rather than answer the items in sequence. They changed their answers only a few times, much less than the high-ability group.

5.6.2 Analysis of Test Reflection Survey questions 1 and 2

Responses on the Test Reflection Survey, described in chapter 4, section 4.3.4, were used to assess perceptions about academic English. The first two questions on the Test Reflection Survey (Items 1 and 2) collected information about the nature of the language in the test. Question 1 prompted test-takers to indicate if (1) yes, they were thinking about academic English while they took the test; (2) no, they were not thinking about academic English as they took the test; or (3) they did not know. The second question asked test-takers if they thought the text in the test was similar to academic English in university textbooks. These responses were collected to elicit perceptions by the test-takers about the content of the test which might reveal if they were thinking about academic English as they responded to the items on the test. If this is true, one would expect a higher percentage of participants in the high-ability group to respond that they were, indeed, thinking about academic English as they took the test and also recognize and respond that the text on the test was similar to the academic language in university textbooks. Responses to these two questions are shown in Figures 5.14 and 5.15, respectively. A chi-square test of independence was performed with the raw counts to examine the relation among the responses by group for both questions.

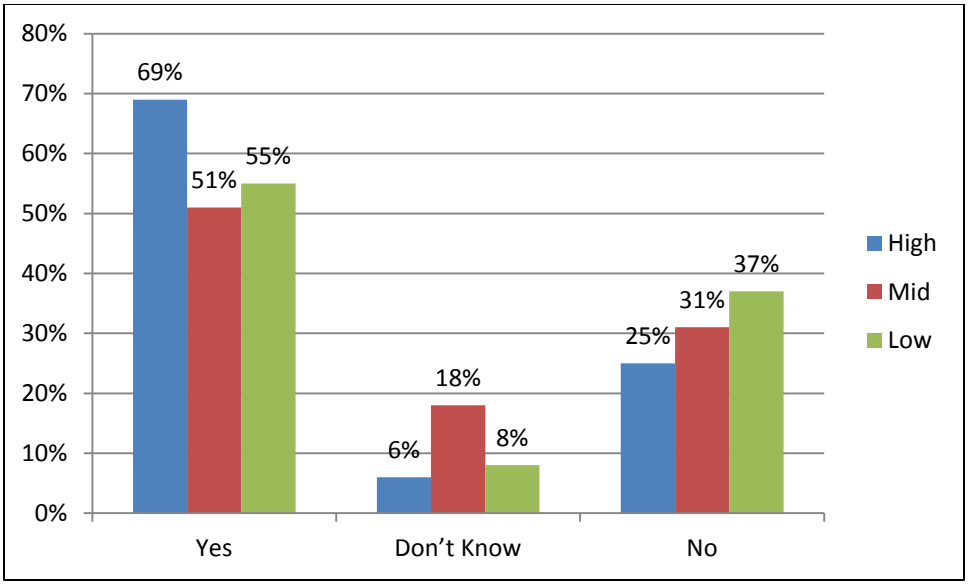

Figure 5.14. Responses about thinking in academic English while taking the test.

According to the responses to the first question about thinking in academic English while taking the test, a larger percentage of test-takers in all groups claimed to be thinking in academic English. However, the percentage of participants that reported that they were thinking in academic English while taking the test did not differ significantly by group, χ^2 $(4, N = 206) = 7.488, p = .112$.

Responses to the second question revealed larger differences among the groups. The percentage of responses indicating that the language was different was largest for the high-ability group (54%), followed by the mid-ability group (39%), and finally, the low-ability group (32%).

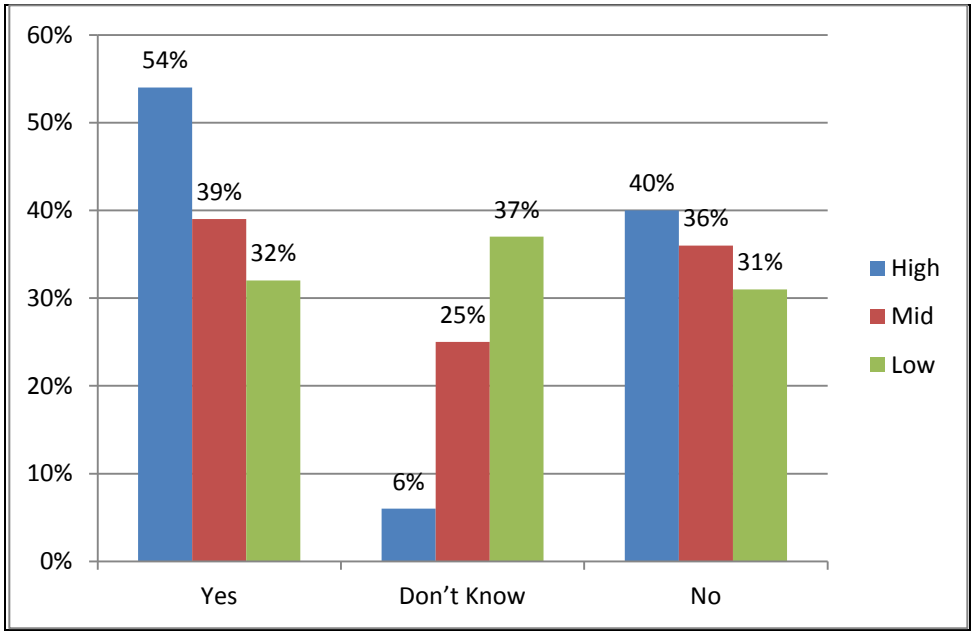

Figure 5.15. Responses about language on the test and in university textbooks

An interesting characteristic of the responses to this question are the increasing percentage of responses to the "don't know" category. These responses theoretically belong to a negative response, because a similarity could not be found between the language on the test and in university texts. Not surprisingly, a significant difference was found, χ^2 (4, N = 196) = 9.711, p = .046*, among the responses from the groups for the second question about whether the language in the test was similar to academic English used in university textbooks. The findings indicate that the groups were different regarding their perception of academic language in the test and in university texts. These results may be interpreted as indicating that some participants may have felt as though they have an idea about characteristics of academic English; yet, this understanding may or may not have played a role as they took the test. Because the results were found to be significantly different for

Question 2, a further analysis of the perception of academic English on the test and in university textbooks was done using responses from the third question on the survey.

5.6.3 Test Reflection Survey data: Question 3

A chi-squared test found significant differences among groups for the responses to Question 2 regarding their perception about similarities or differences between English on the test and in university textbooks. The results from Question 2 were based on a multiple-choice response. Responses to Question 3 on the Test Reflection Survey help explore the differences that were found in the second question of the survey. A Chinese translation was provided to encourage a response that might have been missed due to not understanding the question. The translation was, “请解释一下该考试的题目与大学课本中学术英语应用的相似性或者不同性.” The responses for each test-taker were coded according to the reference made to specific characteristics of academic language. The results from this analysis with test-takers who were able to compare the language on the tests and language in university textbooks are presented in Figure 5.16. Because the number for participants varied in each group, results are shown as percentages.

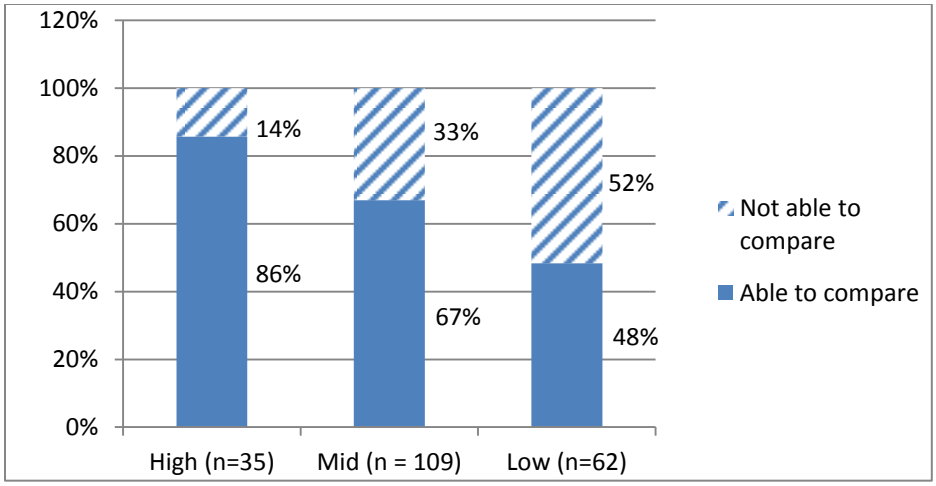

Figure 5.16. Results indicating an ability to compare the texts.

Overall, the high-ability group provided the highest percentage (86%) of responses expressing similarities or differences between the text on the test and text in university textbooks. The mid-ability group provided a lower percentage (67%) of responses showing their group's ability to compare the two texts. The low-ability group responded with the lowest percentage (48%) of responses for their group that contained specific distinctions between the texts. References to similarities or differences between the two texts might be interpreted as showing that the test-takers were familiar with academic English well enough to respond accordingly to this question and indicating that the high-ability group was familiar enough to express the similarities or differences in words.

The data in Figure 5.17 represent the percentage of test-takers in the high-ability group (68%), mid-ability group (67), and low-ability group (48%) who were able to compare the English on the test and the English in university textbooks. The data show three responses for these test-takers.

The distribution is fairly similar among the groups. The majority of test-takers felt there were more differences between the texts than similarities. The largest percentage of responses indicating both differences and similarities were found in the mid-ability group.

Further review of the responses revealed differences that hint at the degree of familiarity with academic English among the responses for each group. Responses in the high- and mid-ability groups, for example, made a connection with academic disciplines. The connections indicated that the English on the test is similar to or different from language found in journal articles.

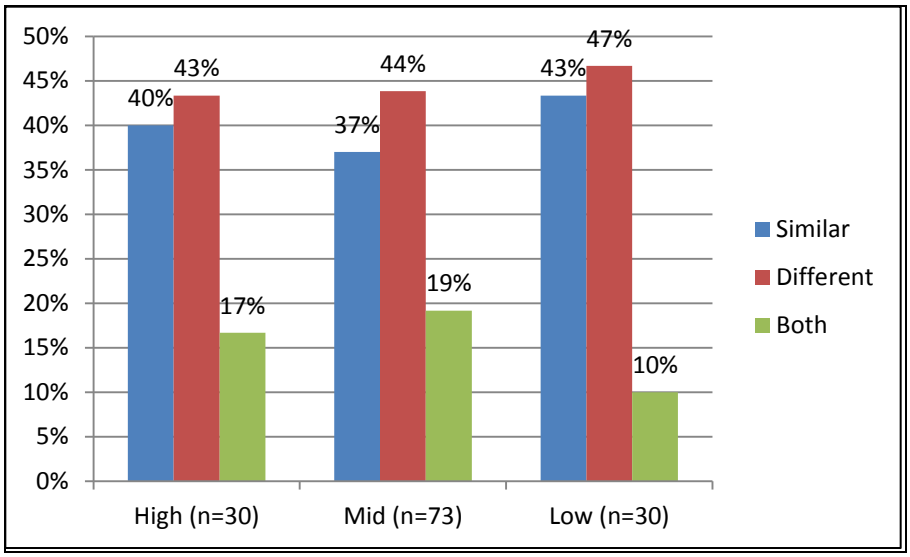

Figure 5.17. Percentage of distinctions indicating similar, different, or both for each group

Test-takers in these groups also thought that the language on the test was related to an academic discipline but perhaps too focused on one or two particular disciplines, as indicated in the following responses.

> *Most of the sentences seem to come from journal articles.* (high-ability)
>
> *It's seems not so professional as the specific subjects in university textbooks.* (mid-ability)
>
> *It is more discipline-specific as it involves a lot of expressions and phrases in the law field.* (high-ability)
>
> *I study Math and Chemistry, so this test is more complicated than my textbook. It's more like languages used in Fiance or economy.* (mid-ability)

Another characteristic of the responses in these groups is the comparison with vocabulary and sentence structure with which the test-taker may have been familiar. Reference is made to sentence structure, passive sentences, rare vocabulary, and complex grammar.

> *In academic English, there will be a lot of terms to be defined. The sentence structure is kind of complex.* (high-ability)
>
> *In general, they are very similar. The most obvious evidence is that the English in this test used a lot a passive sentence.* (high-ability)
>
> *I my opinion, English used in the test is similar to that in the university test books. First of all, this kind of English always use completed sentences instead of simple ones, the formats are different. Second, the words used in academic English are rare than those we use normally. Finally, the conten of the sentence are academic, and the grammar used is complex.* (mid-ability)

Responses in the high- and mid-ability groups indicated different ideas about the content and structure of English on the tests and in university texts. I interpret these comments as an indication that the test-takers are familiar enough with university texts to make an appropriate comment about the distinction between English on the test and English in university textbooks. Overall, the primary understanding of academic English in university

texts by these test-takers might have been that language is used to communicate content in university textbooks. This language is also used by various discourse communities or within a discipline to facilitate communication about a particular topic or subject matter for others in the discourse community.

A number of test-takers in the mid-ability group, however, responded as if their notion of "university texts" was books for teaching English. These may have been university textbooks that the test-takers were familiar with during their EFL or ESL language course at the university. These example responses, not found in the other two groups, show that these test-takers probably had an instructional English textbook in mind when responding.

> *English in text is academic, it is different English used in university textbooks. English used in textbooks teach students how to use English words in life, and how to write essay. English text just text students know some word.* (mid-ability)

> *The similarity is cleared that both of them focus on grammar and using for the verb. It teach students how to use the word correctly have a better understanding of the word. It improves the skill of written English by checking the correct answers. However, most of the University text book, the test maker only focus on the exact answers. It will create fixed questions and made only one correct answers. In comparison, this test is more open and provide variety of choices for students. It help students to create more answers and think out more verbs to use. The method help students to have a better understanding on the written English.* (mid-ability)

Finally, the low-ability group had the lowest percentage of responses that explained similarities or differences between the text on the test and text in university textbooks (48%). Many of the responses were general in nature and did not address specific characteristics of the texts. Some responses from this group addressed the purpose of the language on the test and in academic texts. According to these test-takers, university texts are used to convey content, whereas the English on the test was used to assess language.

> *Similarity: most of words are not simple, oppositely, it represents that very formal in the academic English. Difference: 1.University textbooks: it more focus on students to*

> *understand the content so that it would not use too hard words. 2. Test: it is a examination to test student's capability for applying words and grammars; thus, the purpose is not same to compare with university textbooks.* (low-ability)

> *In text book, they almost have complete sentences, and we do not need to find which word I need to put it into a sentences. I think the aim at this test is let teacher know how many words you have already know, and how to use them. However, I think it may mislead the meaning of the sentences. And In Text book. they want us know the right ways and knowledge of something. So I think it is completely different compared the English in this test and English used in university textbooks.* (low-ability)

Additional responses from this group might be an indication of an awareness of academic text but did not give a reason how they are similar or different, as in the following response:

> *It is parity similar with the academic English which be used in university text book. Because both of them are difficult to spelling and I hardly know them.* (low-ability)

This response indicates that the test-taker does not have a command of academic language ability and considers that any language that is difficult might be academic language.

Responses in this group also indicated a text book example of what academic English should be rather than providing an understanding of features and characteristics of the language. This response from a low-ability test-taker, for example, might refer to the discourse features of an academic text that was taught to the student in an English class.

> *In my opinion, the university textbooks should have a introduction and a conclusion, but this passage has not.* (low-ability)

The idea of academic language to this test-taker was based only on the organization of the text at a discourse level and not vocabulary or content at a sentence level. Organization was difficult for this test-taker to identify in a test with only sentence level items. Interestingly, a few responses that did not explain a similarity or difference did reply with an explanation but only stated that the test-taker was not familiar with academic English. Responses such as these were not found among the mid- or high-ability test-takers.

> *I will start my university in some months, I do not know about English used in university, so I am not be able to explain it, or I would not be duty.* (low-ability)
>
> *I don't know what English used in university textbooks because I haven't read any university text book yet. I am sorry that I can't give you any answer to this question.* (low-ability)
>
> *I don't know.* (low-ability)

In sum, the group with the highest percentage of responses explaining the differences between the language on the test and language in university texts, the high-ability group, provided more specific examples for similarities and differences. Their responses were similar to responses from the mid-ability test-takers with reference to academic disciplines, vocabulary, and sentence structure, which could indicate a familiarity with the academic language used in university texts. One characteristic that was unique to the mid-ability test-takers was the interpretation of academic texts as instructional materials.

Responses in the low-ability group were mixed. A number of responses related to the purpose of the texts but were not specific about sentence level characteristics such as vocabulary or sentence complexity. A few responses commented on discourse level characteristics that might have been presented as characteristic of academic language taught in an English writing course. Finally, many responses from the low-ability group were "I don't know." At least test-takers in this group were honest in many responses stating that they were not familiar with university texts.

5.6.4 Post-test interviews

Additional qualitative data were collected to help explain the quantitative results. Post-test interviews were conducted with five test-takers from three groups. The structure of

the semi-structured interview can be found in chapter 4, section 4.3.6. The objective was to interview a test-taker with a high score and a test-taker with a low score on the Collocational Ability Test from each group. The interviewees' ID number, their group affiliation, and scores on the Collocational Ability Test using both the dichotomous scale and the partial credit scale are presented in Table 5.10.

Table 5.10. Interviewees who participated in the post-interview

ID#	Group	CAT score	CAT score PCS
569	High	18	24.5
1094	High	3	14
940	Mid	13	20.5
700	Mid	2	9.5
825	Low	11	16.5
891	Low	2	7.5

Although the sample was small, the responses were fairly consistent with higher performing interviewees versus lower performing interviewees. A few trends started to emerge from the responses.

The aim of the first question was to identify if the interviewee were aware of the purpose of the test. Interviewees were asked if they understood what the test was testing. Comments from interviewees in response to this question are show below in Table 5.11. The interviewees who performed better on the test from all three groups were aware that the test was measuring some type of vocabulary or phrase, "words you can put together" (ID 569), and interviewee ID825 even added vocabulary comprehension "in context." The term *phrase*

was an indication that the interviewees were aware of the idea that certain words are frequently learned together as a unit.

Table 5.11. Comments to the question about the purpose of the test.

ID #	Response
High-performing interviewees	
569	*"Vocabulary… Phrase… What kind of words you can put together in academic language"*
940	*"Phrases and words"*
825	*"Vocabulary comprehension of content"*
Low-performing interviewees	
1094	*"I don't know"*
700	*"Vocabulary. Some good words to match the sentence"*
891	*"The verb. English test just"*

The responses from the lower performing interviewees were less precise, ranging from not knowing to a simple "vocabulary, some good words to match the sentence," which could mean single lexical items without attention to lexical relationships such as restricted collocation. None of the interviewees in either category provided the term *collocation* in their response.

The second part of the interview investigated the understanding of academic language, the process of thinking about academic language during the test administration, and any evidence regarding reading difficulty. Responses to this question are shown in Table 5.12.

Table 5.12. Comments to the question about academic language

ID #	Response
High-performing interviewees	
569	*"Yes"*
940	*"I know what academic English is."*
825	n/a
Low-performing interviewees	
1094	*"No, is there a difference? If I can think of the academic word, I would use it."*
700	*"Mostly generally English."*
891	*"I just think it's an English test."*

The question regarding whether the interviewee was thinking in academic English was not easy to answer because it relies on an understanding of what academic English is. While the high-performing interviewees may indicate they were thinking in academic English, they may be answering in the affirmative about what they think academic English is.

More convincing evidence comes from clear responses from the low-performing interviewees. In response to the question of whether he/she was thinking in academic English while taking the test, interviewee ID1094 replied, "No, is there a difference?" The reply by the other low-performing interviewee was, "Mostly general English. If I can think of the academic word, I would use it." These responses showing a lack of knowledge about academic language are more revealing than a positive response from the high-performing interviewees.

The next question probed whether reading difficulty may have been an influence. The interviewees were asked if they had difficulty understanding any of the sentences. Responses to this question are provided in Table 5.13.

Table 5.13. Comments to the question about reading difficulty

ID #	Response
High-performing interviewees	
569	*"Sometimes I didn't know understand the whole sentence {} there is a subject and a verb and you make that complex sentence into a shorter one and combined with your knowledge of those combinations. That's what I got ….[] my test."*
940	This interviewee could understand most sentences. Reading did not seem to impede understanding or be a problem.
825	*"Sometimes it's difficult to understand some of the sentences in the test."*
Low-performing interviewees	
1094	*"Yes, some words are new words for me."*
700	n/a
891	*"I'm still not accustomed to use English to become my daily language."*

Most interviewees, regardless of level, indicated some difficulty with reading comprehension. Interviewee ID569, who had the highest score, thought that some of the questions were potentially culturally related and were more difficult for an "international student" to understand. Item 17 was identified as problematic by this interviewee.

> ID569: *I'm not sure if #17 is culturally related but I don't' think I understand it clearly*
> Researcher: *even though you didn't understand this one, you got it correct.*
> ID569: *Yes, because "take action" is a fixed phrase.*

This interviewee (ID 569) explained her approach to the test when she did not understand the whole sentence. She would look for the subject and the verb in a complex sentence to make a shorter sentence. This strategy, combined with her knowledge of the word combinations, "fixed phrases," assisted in providing a correct target response, even if the meaning of the entire sentence was not comprehensible. An interesting comment by a number of

interviewees reflected a conscious effort to translate each sentence into Chinese to understand the meaning. Although no interviewee admitted to direct L1 transfer, it might have occurred subconsciously.

The next question uncovered information about prior experience regarding learning, memorizing, or being exposed to verb-noun collocations in class or English study. Responses can be seen in Table 5.14.

Table 5.14. Comments to the question about learning collocations

ID #	Response
High-performing interviewees	
569	*"Only in the English text book. The English teacher will give you a list of fixed phrase that you must pay attention to."*
940	*"Studies most common phrases in English class. I remember very beginning was just phrases."*
825	*"I have studied word pairs V_N and phrasal verbs in class."* *"I have memorized lists of word pairs. Although they are easier to remember in a sentence."*
Low-performing interviewees	
1094	*"We used to study lists of word to prepare for tests but didn't remember very well or very fast."*
700	*"I can't remember. Maybe someone will mention it in one or two seconds."*
891	n/a

A trend appeared regarding the use of collocation lists in language learning. All of the high-performing interviewees clearly remember using lists of "phrases" or individual words. The lists were decontextualized lists that students were expected to memorize. The low-performing students remarked that they were aware of lists but did not memorize them. One

interviewee (ID1094) recounted starting with a list and learning all the A and B words but stopping at letter C because there were too many words.

Interviewees also commented on their responses to particular items on the test. The item with the highest item facility was Item 26. Interviewees were given the opportunity to see the target response and their own response and comment on what they were thinking as they responded to the item. Responses by the interviewees, credit given, and their interview comments are shown in Table 5.15.

Item 26

Particular attention will be _____ to differences in political culture and the role of linkages between local and regional interest groups in political developments.

Table 5.15. Comments about Item 26, "pay attention"

ID #	Response (Credit)	Interview comments
High-performing interviewees		
569	paid (full)	***"Pay attention to** is we call it fixed phrase"*
940	payed (full)	*"**Pay attention to** is a phrase"*
825	Paid (full)	*"Pay attention to"*
Low-performing interviewees		
1094	noticed (none)	*"**Pay attention** I know." "Attention and notice are related"*
700	put (partial)	No comment
891	send (none)	This interviewee thought that culture was the node in the collocation.

It is clear that the performing interviewees were all familiar with the collocation "pay attention" as a phrase including the preposition "to." One of the interviewees (ID1094) in the

lower performing group admitted knowing the word pair "pay attention." This comment was made after looking at the target response. Her next comment sheds light on how she approached her response. Rather than providing a verb as a collocate without the meaning reduced due to the combination of the pair, she was making an association with a verb and noun in the sentence based on a similar meaning. This interviewee was equating her response "notice" with the node "attention" as having a similar meaning. Noticing something and paying attention to something have similar meanings. The other lower performing interviewees were either not able to express a reason for their answer or had chosen the wrong node in the sentence.

Item 28 was the most difficult item on the test. Again, interviewees commented on their responses to this item. These comments are shown in Table 5.16. None of the interviewees in either group were awarded full credit for this item. The higher level interviewees reflected on their responses, revealing possible blending, as suggested by Howarth (1996), as a reason for producing a particular response. This occurs when the interviewee produces a verb from a collocation that is familiar but is mistaken to be an acceptable combination with a node in a different context. This seems to be the case with interviewee (ID825), who indicated that she was thinking of "get an idea" and generalized this combination to getting something else such as "attention." Interviewee (ID569) also indicated that she was thinking of a different phrase or collocation when responding to this item. Both instances can be considered relevant to uncovering partial knowledge of collocations known by the interviewees.

Item 28

The motivations - both many and complex - behind Japan's desire to exert her influence in Asia have _____ considerable attention from historians.

Table 5.16. Comments about Item 28 "receive attention"

ID #	Response (Credit)	Interview comments
High-performing interviewees		
569	taken (partial)	*Yeah, take consideration. Take considerable attention. Oh, I think I made a mistake. I thought it was* ***take into consideration.***
940	caused (none)	*I didn't pay attention to the last two words* ***from historians****.*
825	Got (partial)	Never heard "receive attention" Never heard "get attention" *It's similar to* ***get an idea****.*
Low-performing interviewees		
1094	taken (partial)	Chose past participle because of grammar. *"No confidence because I just guess."*
700	made (partial)	Researcher: "*Did you understand that sentence?"* 700*: "I understand this sentence because I know the words. Maybe I only can think some words and made is the best choice. Maybe if you put receive as a choice I could choose it. Can we say 'make attention'?"* Researcher: *[COCA search] "Only one time in academic texts."* 700*: "I've only heard pay attention. Laughs. Receive attention."*
891	make (partial)	*"If I don't know, I use words like: make, get, do."*

All three interviewees in the low performing group admitted to guessing for Item 28. A common theme within this group was the use of verbs such as *make*, *take*, and *do* when they were unsure of the target collocation. Unfortunately, because these verbs are so common, partial credit was given using the partial credit scoring method if one of the verbs

was a commutable verb. In these cases, credit that should be awarded for partial knowledge was actually rewarding guessing, when no relationship between lexical items was justified. A rebuttal for this claim might be that the interviewee had a choice of a variety of highly frequent verbs—*make*, *take*, *get*, *do*, and *have*—and did select one that was a commutable verb, so perhaps there is unconscious partial knowledge on some level.

The answer to RQ 3 comes from responses to the Test Reflection Survey and post-test interviews with a sample of interviewees. The perception of academic collocations and academic language varied among the groups. The extent to which interviewees were actually thinking in academic English as they took the test is unclear. The responses to questions 2 and 3 on the Test Reflection Survey indicated, however, that as proficiency level increases, a clearer understanding of the meaning of academic language is achieved. The high-ability group described characteristics of the academic language they were familiar with as a comparison with the text on the test. The mid-ability group responded with comments that indicated a lack of agreement about the content of academic textbooks. Some interviewees commented on specific disciplines, whereas other indicated that the textbooks they had in mind were used for English instruction.

Other differences existed among the groups regarding the purpose of the test and experience learning collocations. The higher performing interviewees tended to understand the test was measuring collocational ability or some type of phrasal ability. The lower performing interviewees either did not know or said that it was a vocabulary test with a focus on individual words. These responses correspond with the experience learning English for the two levels. All higher performing interviewees recalled learning lists of collocations or phrases in their English instruction. Most of the lists did not include a context. Emphasis was

on rote memorization. The lower performing interviewees either did not remember any such lists or admitted that they did not take the time to memorize the lists.

Finally, the quantitative and qualitative data regarding students' perception of academic collocations and academic language appear to match the performance on the Collocational Ability Test. It also corresponds with the observations about metacognitive and cognitive strategy use while taking the test. This correspondence can be seen as evidence to support the assumption that interviewees use metacognitive and cognitive strategies that are related to collocation use in academic language while taking the test. This evidence supports the warrant that expected scores are attributed to a construct of collocational ability in academic writing, thereby further supporting the explanation inference.

5.7 Group differences

RQ 6 was developed to add support or challenge the assumption that test-takers at different proficiency levels perform differently on the collocation test based on their proficiency level. This study included three general proficiency levels: students in an intensive English program, undergraduate students who still require instruction in academic language, and graduate assistants who have been using academic language for a period of time. The results of this study would be expected to show that test-takers who belong to the high-ability group outperform the other two groups. This group would also be expected to be significantly different because they are a sample of students who are engaging in academic language. Sample sizes for all groups were large enough using both scoring methods to use a one-way independent ANOVA to test for differences. Statistically significant differences among these groups could be seen as evidence to support the warrant that the construct of

collocational ability as assessed by the Collocation Ability Test accounts for the academic collocational language performance in academic discourse in English-medium colleges and universities. Backing for this warrant would support the extrapolation inference in the interpretive argument.

5.7.1 Distinctions among proficiency level groups using dichotomous scoring scale

The box and whisker diagram in Figure 5.18 shows score distributions for three proficiency levels on the collocational ability test using the dichotomous scoring scale. The three groups are high-ability participants (n = 35), mid-ability students (n = 109), and low-ability students (n = 62).

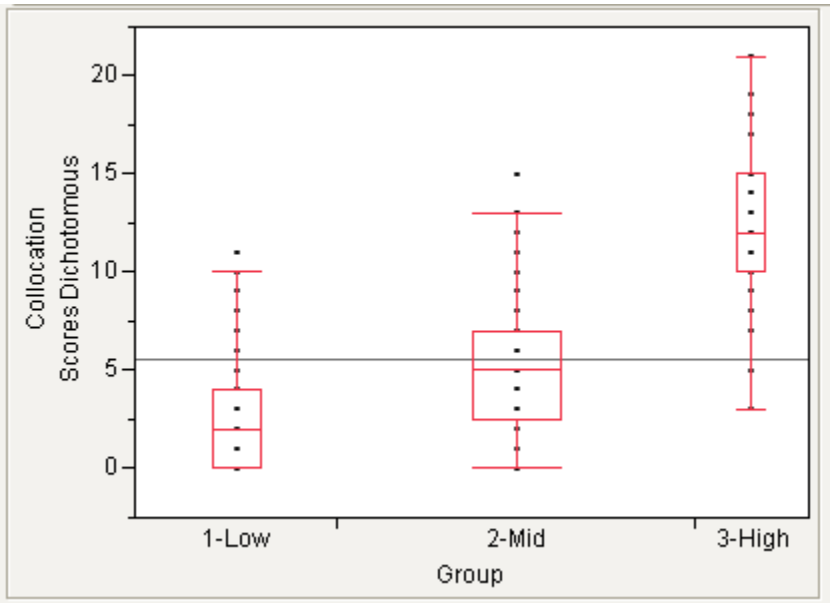

Figure 5.18. Box and whisker diagram showing score distributions for three proficiency levels on the Collocational Ability Test

Using JMP 9 statistical software, a one-way independent ANOVA was used to test for differences among the three groups. Significant differences among groups were found on scores on the collocation tests, $F(2, 203) = 91.93$, $p < .0001$. A Tukey HSD Post-Hoc test indicated that the groups were significantly different from one another ($p < .01$).

5.7.2 Distinctions among proficiency level groups using polytomous scoring

Score distributions are presented in Figure 5.19 as a box and whisker diagram for the same three groups using the polytomous scoring method. Results from a one-way ANOVA showed that the three groups were significantly different, $F(2,203) = 64.50$, $p < 0.0001$.

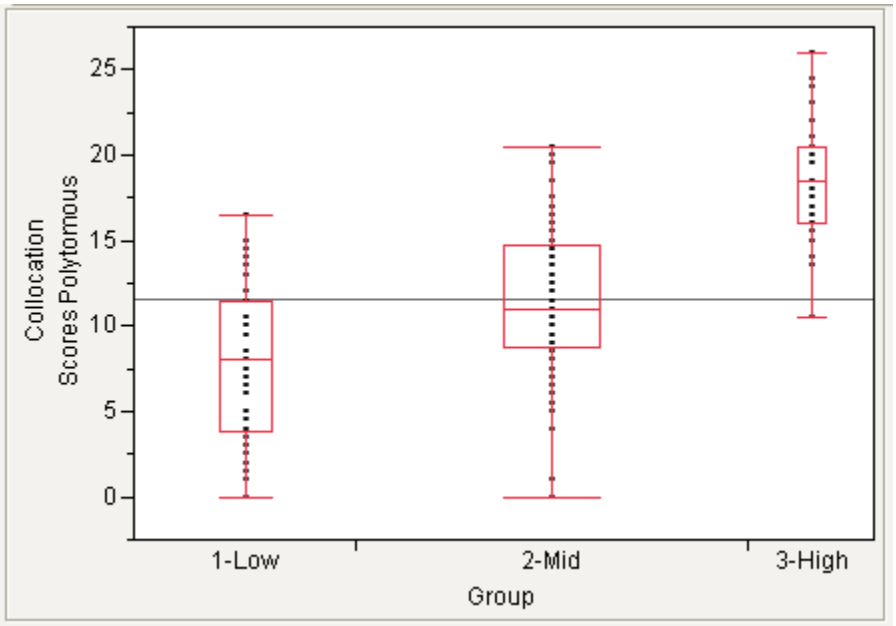

Figure 5.19. Box and whisker diagram showing score distributions for three proficiency levels on the Collocational Ability Test using the polytomous scale

A post-hoc test was used since the ANOVA showed a significant difference among the three groups. A Tukey HSD post-hoc test showed that all three groups were significantly

different at $p < .01$. Distinctions were found among the three groups using both scoring methods. The research assistants performed best, followed by the undergraduate students, who are enrolled in content classes and English language classes that focus on academic language. Both of these groups of test-takers who are currently using academic language outperformed the students in the intensive English program, who may not have a large amount of knowledge or exposure to academic language.

The results from these tests answer the research question regarding the performance on the collocation test based on proficiency level. Significant differences support the assumption that test-takers in higher proficiency groups will perform better because of the greater exposure, experience, and knowledge of collocations in academic language. This provides backing for the warrant that the construct of collocational ability as assessed by the Collocational Ability Test accounts for the academic collocational language performance in academic discourse in English-medium institutions. Support for the assumptions underlying this warrant provides backing for the extrapolation inference.

5.8 Academic language

Results helping to answer RQ 7 are intended to be used to support or challenge the assumption that there is a relationship between the construct of academic collocations and the construct of academic vocabulary identified as single lexical items. A positive correlation between scores on the Collocational Ability Test and another test of academic vocabulary would support the assumption that a positive relationship exists between these two measures. Correlations for scores of both scoring methods on the collocational ability test and subsections of vocabulary size test for all test-takers are presented below in Table 5.17.

Measures in italics in Table 5.17 indicate non-normal distributions. Spearman's rho was used to calculate the correlation coefficient that included any of these non-normal data sets. The parametric Pearson correlation was used with two data sets with normal distributions. In order to account for the measurement error in the estimates, the correlation coefficients were corrected for attenuation.

Table 5.17. Correlations for scores of both scoring methods on the collocational ability test and subsections of vocabulary size test for all test-takers (N=206)

	collocation test scores (dich) k = 35 alpha=.89	collocation test scores (poly) k=35 alpha=.83	*vocabulary test scores* k=30 alpha=.87	*vocabulary 2k items* k=10 alpha=.71	*vocabulary content items* k=10 alpha=.68	vocabulary academic k=10 alpha=.77
collocation test scores (dich)		1.00	.74[1]	.70[1]	.60[1]	.68[1]
collocation test scores (poly)	1.00	.98[1]	.72[2]	.73[1]	.55[1]	.70[2]

[1] Spearman's Rho correlation significant at the 0.0001 level (1-tailed)
[2] Pearson correlation significant at the 0.0001 level (1-tailed)
All correlations corrected for attenuation.

The reliability estimate for the total scores on the vocabulary size test was high (α = .87). The estimates on the sub-tests were lower as expected with shorter tests. The content words sub-test was the lowest estimate (α = .68), followed by the high-frequency words sub-tests (α = .71). The sub-test with the academic words has the highest reliability estimate (α = .77).

Looking further at the sub-tests on the vocabulary size test, a higher positive relationship would be expected between scores on the Collocational Ability Test and the sub-test of academic vocabulary than measures of general high-frequency vocabulary because the

collocations were sampled from a corpus of written academic language, as the single lexical items and collocation pairs would both be representative of general written academic language. This is not the case, however. The correlational results indicate that the relationship with the dichotomous method was higher for the high-frequency vocabulary sub-test (.74) than for the sub-test with academic vocabulary (.68). Correlations using the polytomous scoring methods indicate a similar relationship with the two sub-tests; the relationship with high-frequency vocabulary was higher (.72) than with academic vocabulary (.70).

The polytomous method resulted in slightly higher correlations than the dichotomous method with the high-frequency and academic sub-tests. There was an inverse relationship with the sub-test developed using content vocabulary from the collocation test. The correlations with the content subsection were higher for the dichotomous scale (r_s = .60) than for the polytomous scale (r_s = .55). A low correlation on the content section of the vocabulary size test may indicate that some of the lexical items that make up the word combination pairs were possibly either unknown or partially known to the test-taker. If this is true, this might suggest that the word combination was unfamiliar to the test-takers, which inhibited their production of a collocation as a single lexical item. Evidence from this analysis supporting the assumption underlying the extrapolation inference was weak. Nonetheless, the polytomous scoring method did result in higher correlations overall than the dichotomous scoring methods.

5.9 Student placement

Scores on the Collocational Ability Test are intended to be used to contribute to the decision to guide their placement in English-medium instruction without any ESL courses, in

English medium instruction with an appropriate level of ESL course, or in an appropriate English language course. Ideally, this use would be supported by findings that show that scores on the Collocational Ability Test would place a targeted sample of test-takers correctly. To produce such findings, one would need a sample of test-takers who were known to be placed correctly. The sample of test-takers used in this study cannot be assumed to be placed completely accurately, but the three subgroups in the sample are expected to be represent different levels with respect to the targeted language construct, and therefore it is informative to see how the Collocational Ability Test scores of these groups would place test-takers into the three groups of which they are members.

Cut-scores were determined using the final steps of Zieky and Livingston's contrasting groups method (see Brown, 2005; Fulcher & Davidson, 2007). Distributions of scores from the Collocational Ability Test using the polytomous method for all groups were compared visually (see Figure 5.20). Trendlines were used to display the distributions graphically, producing smooth curves that overlap. Areas where the distributions overlap were examined and cut-scores were set at the point where two distributions intersect. These are the points that separate the "non-masters" from the "masters." The lower cut-score was set between the distributions for the low- and the mid-ability groups. The higher cut-score was set at the point where the distributions intersected for the mid- and high-ability groups. Cut-scores thus were set at 13 for placement into English-medium courses with additional English instruction and 37 for matriculation without the need for additional English instruction.

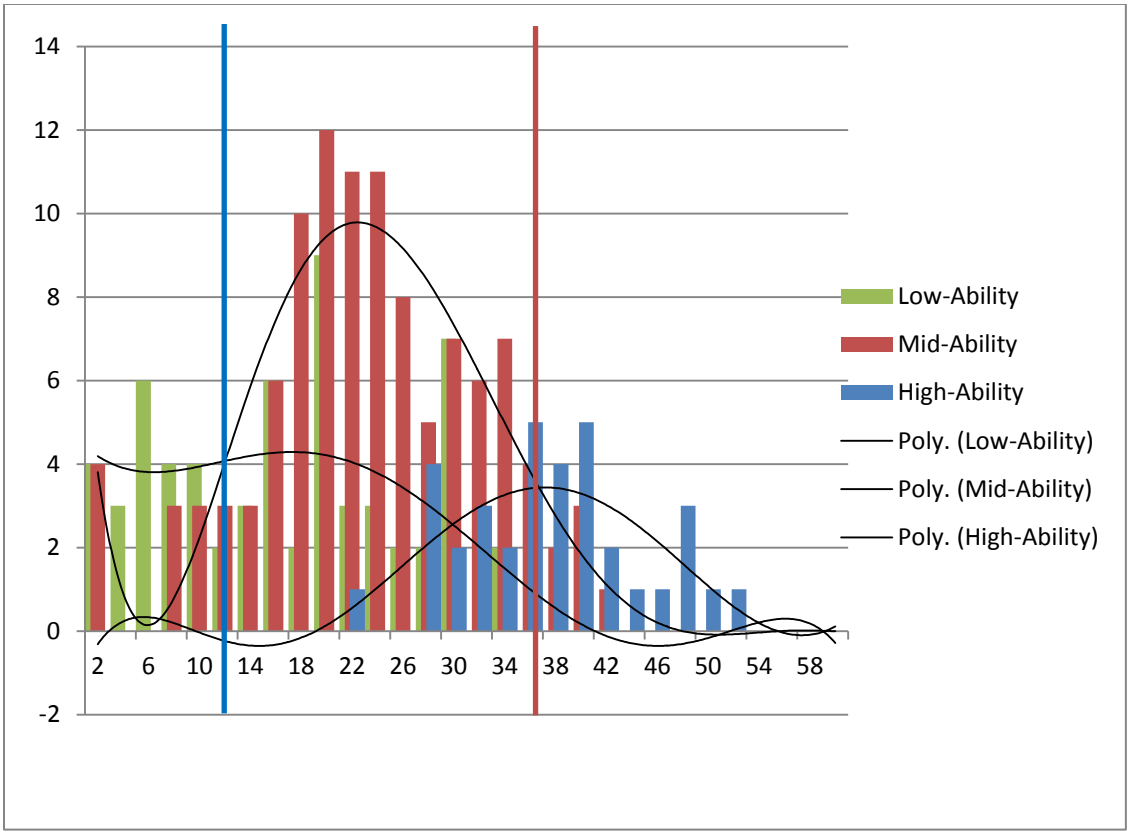

Figure 5.20. Overlapping of the three ability-level groups' distributions on the Collocational Ability Test with trendlines

The data in Table 5.18 show the number of test-takers in each group placed by these cut-scores. Test-takers in the low ability group would be expected to obtain scores below the first cut-score of 13. Their performance on the test should indicate that they are not ready to be enrolled in an English-medium college or university and need to be placed into ESL courses. The mid-ability group in the study should obtain scores between and including the two cut-scores (13-37), which would indicate that test-takers were ready for university study but would benefit from additional language instruction while enrolled in the college or

university. Test-takers in the high-ability group would also be expected to have scores above 37. They would be ready to matriculate without additional English language courses.

The analysis shown in Table 5.18 reveals that the expected trends occurred for the mid- and high-ability groups when the placement based on a priori groupings was compared to those based on the cut scores. The largest number of test-takers in each band corresponds with the group placement by proficiency for the high- and mid-ability groups. In contrast, there was considerably less correspondence for the low-ability group between a priori grouping and placement based on scores from the Collocational Ability Test, with 63% of the test-takers misplaced. The data also point out that use of the Collocational Ability Test with these cut scores would result in 76 (37%) misplaced test-takers in all three groups.

Table 5.18. Placement of test-takers by cut-scores

Group	0-12	13-36	37^{+}	Potentially misplaced	% Potentially misplaced	Total
High-ability	0	18	19	18	51%	35
Mid-ability	13	90	6	19	17%	109
Low-ability	23	39	0	39	63%	62
Total				76	37%	206

Fifty-one percent of the test-takers in the high-ability group would be potentially misplaced if placement decisions were based only on scores from the Collocational Ability Test. Approximately half of the test-takers in this group (n = 19) would be allowed to matriculate to an English-medium college or university without the need for additional English language instruction, based on the upper cut-score of 37. The other test-takers in this group (n =18), however, would be allowed to matriculate but would also be placed in English

language courses to improve their academic English skills, if this test were the only source of decisions about placement. None of the test-takers in the high-ability group would be placed into ESL alone.

Seventeen percent of the test-takers in the mid-ability group would be potentially misplaced. A majority of test-takers in the mid-ability group (n = 90) would be placed correctly in English language courses while enrolled in credit-bearing courses, following the lower cut-score of 13. Again, using only the scores on the Collocational Ability Test for placement, a few test-takers in this group (n = 6) who scored 37 or higher would be allowed to study without additional English language instruction, whereas the rest (n = 13) would not have been admitted to the college or university.

Sixty-three percent of the test-takers in the low-ability group would potentially be misplaced. No test-taker in the low-ability group would be allowed to matriculate without additional language instruction following the upper cut-score. Placement based on the lower cut-score would allow 39 test-takers to study at an English-medium institution with additional English instruction. The remaining 23 test-takers in the low-ability group would not be allowed to matriculate with or without additional English instruction.

Overall, the scores on the Collocational Ability Tests would place a number of test-takers in each ability group following test-takers' current placement. If, however, placement decisions were based only on scores from the Collocation Ability Test using the cut-scores determined by overlapping distributions, 37% of all test-takers would be placed differently. The mid-ability group has the fewest percentage of potentially misplaced test-takers (n = 17%) followed by the high-ability group (n = 51%) and the low-ability group (n = 63%). These results indicate that the scores on the Collocational Ability Test do not discriminate

with precision among the three groups of test-takers, but the test-based placements are consistent with the original groupings. . It remains to explore how placements informed by both the Collocational Ability Test and other placement information might distinguish more precisely among the three groups.

5.10 Chapter summary

This chapter presented the results obtained from the quantitative and qualitative data collocations. The data were used to answer the research questions presented in chapter 2. Each research question was developed based on an assumption underlying a warrant supporting an inference in the interpretive argument. Qualitative results from screen capturing, responses on the Test Reflection Survey, and comments during post-interviews were used to support and confirm the results from the qualitative data collected from the Collocational Ability Test, reading test, and vocabulary test. The overall findings resulted in varying degrees of evidence supporting the assumptions, which provide backing for the warrants underlying the inferences in the interpretive argument. An exploratory analysis of student placement using only the scores from the Collocational Ability Test indicated that a number of students would be placed differently than their current placement.

The following chapter presents an overview of this study, beginning with test development, scoring methods, instruments, measures, and methodology. The chapter also draws conclusions from the data which were found to answer the research questions as support as a validity argument. The validity argument is an evaluation of the interpretive argument. Theoretical and empirical evidence found in the study are presented in light of support of the validity argument of the Collocational Ability Test. The final chapter also

discusses the limitations of the research and the limitations of the obtained results for the validity argument. The chapter concludes with recommendations for future research.

Chapter 6 Conclusion

The Collocational Ability Test was designed as a measure of academic language which might be used as to contribute to or strengthen decisions allowing a test-taker to matriculate to an English-medium college or university or be placed into an appropriate English language course for instruction in academic language. The test was developed using a corpus-driven approach by identifying target collocations from a target domain corpus of written academic English language beginning with the automatic identification of word pairs in the target language corpus. This was followed by a manual selection of restricted verb-noun collocations. The selected collocations were developed into test items using concordance lines from the corpus.

Two scoring methods were developed to score the responses on the test. The key for the dichotomous scoring method identified a correct response as the verb that was identified as part of the collocation for the item that was identified as a whole unit based on frequency in the initial selection process. All other responses were not given credit. The polytomous scoring method awarded full credit to the target response as identified according to the dichotomous scoring method. Partial credit was also given to responses that were verified in an American English corpus of academic language with five or more occurrences. Responses that did not appear or were found with very low frequency (i.e., less than 5 times in 86 million words) did not receive any credit.

Test-takers also took a reading test and a test of productive vocabulary size. The scores from these tests were compared with the scores on the Collocational Ability Test. A Test Reflection Survey was administered to elicit information about test-takers' perception of academic language on the test and in university textbooks. A post-test interview was also

conducted with a sample of test-takers to identify cognitive and metacognitive strategies used by test-takers during the test. Computer screens were recorded for an additional sample of test-takers for this same purpose of investigating the test-taking process.

This mixed-methods study was intended to support the results from the quantitative data analysis with results from qualitative data collected. The data collected and the analyses from the quantitative and qualitative data provide evidence to support the interpretive argument for the Collocational Ability Test presented in chapter 2. Each inference was based on a warrant that had underlying assumptions that needed to be supported. Support for the assumptions came from theoretical rationales and/or empirical data analysis. A number of assumptions were supported by findings in previous research on collocational knowledge and language assessment. Backing for other assumptions could only be supported by empirical evidence once scores had been collected for the test. These assumptions guided the development of the seven research questions which were answered in chapter 5.

The rest of this chapter presents a summary of the evidence that was found to support or challenge the interpretive argument. An evaluation of the evidence supporting the interpretive argument is presented as a validity argument for the test. Where applicable, findings from both scoring methods are presented. The chapter ends with limitations and implications of the study, suggestions for future research, and a brief conclusion.

6.1 Validity argument

The interpretive argument that expresses the intended score meaning for the Collocational Ability test provided the foundation for the type and amount of evidence needed for a validity argument. By taking the validity argument into consideration early in

the test development process, it can assist in the design as well as evaluation of the test. Such planning has been called the *design validity* (Briggs, 2004).

This study provided logical and empirical evidence as backing for the seven-part interpretive argument for the collocational ability test. The grounds for this interpretive argument began with the domain of academic English use, and the argument was built by providing backing for the assumptions underlying four inferences: evaluation, generalization, explanation, and extrapolation. Backing for the utilization inference and impact intention inference, which are part of the interpretive argument, have not been presented in this study since the test is not fully operational. The use of the test scores and the intended beneficial consequences of the test cannot be evaluated until the test is being used. Where applicable, the results from the dichotomous and the polytomous scoring methods were compared and evaluated to determine which scoring method would prove most beneficial in providing scores that could be useful for the intended score interpretation and purpose of the test. Figure 6.1 shows the types of intended backing collected for each step in the validity argument for the Collocational Ability Test. Figure 6.1 is based on the steps of the TOEFL validity argument (Chapelle et al., 2008, p. 349). Each step is supported by the backing provided so that one can climb the stairs sequentially to get to the top. The intended backing for the utilization and impact intention inferences are shown in parentheses since neither theoretical rationales nor empirical research have been obtained for these steps.

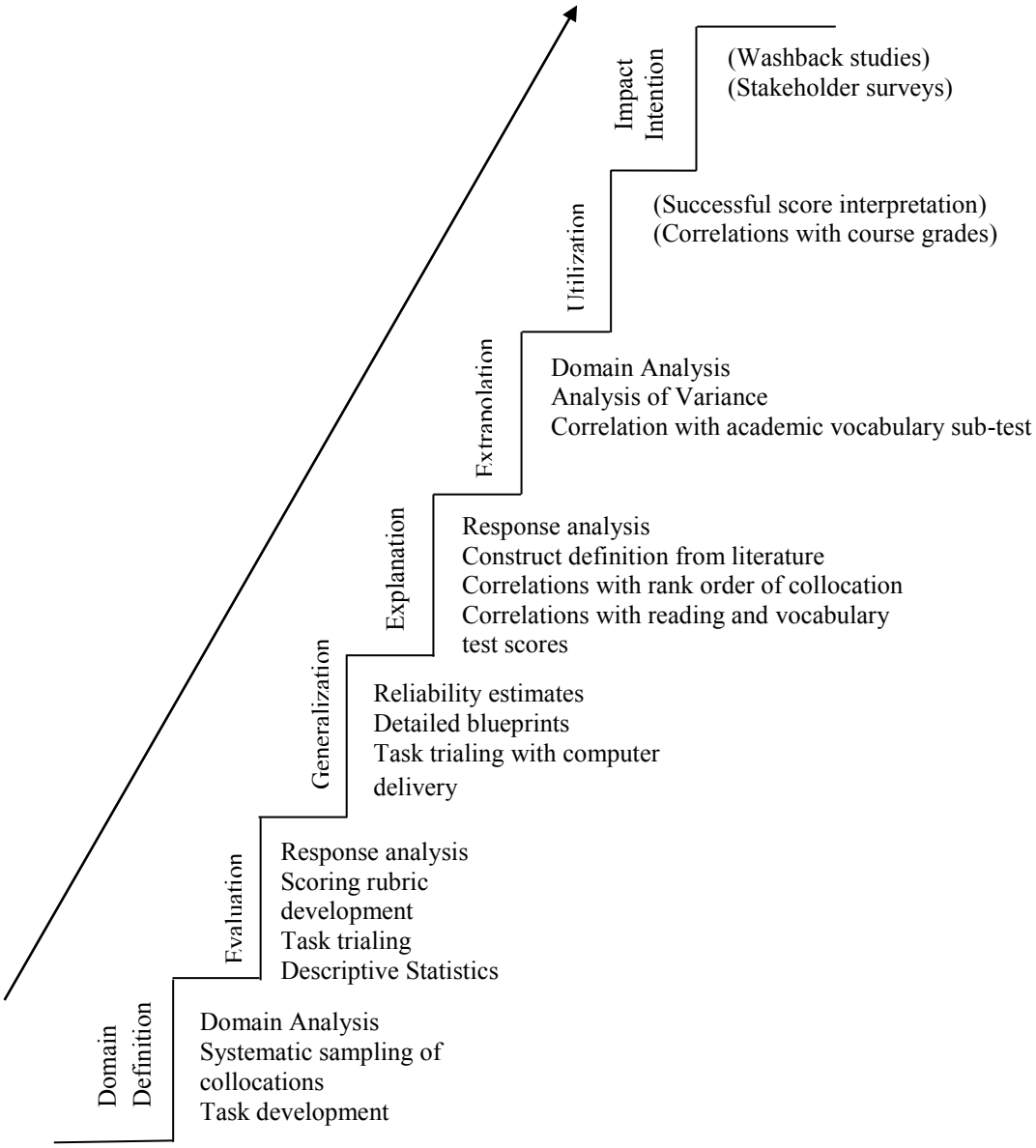

Figure 6.1. Backing collected for each step in the validity argument for the Collocational Ability Test

6.1.1 Domain description inference

The interpretive argument began with the domain description inference, which linked performance in the target domain to the sample of observations on the test. This inference was based on the warrant that observations of performance on the collocation test reflect the collocational ability that is relevant and representative of the target domain of language use in English-medium colleges and universities. The plausibility of the warrant was found in the support of the three assumptions that underlie it. Backing for these assumptions came from an analysis of the target domain, systematic sampling of collocations and task development, which were supported by other studies in the literature. Table 6.1 shows the warrant, assumptions, and types of backing used to support the assumptions for the domain description inference.

Table 6.1. Warrant, assumptions, and types of backing for the domain description inference

Warrant licensing the inference	Assumptions underlying warrant	Types of backing used to support the assumptions
Observations of performance on the collocation test reflect the collocational ability that is relevant and representative of the target domain of language use in English-medium colleges and universities.	1. Performance in written academic English tasks relies in part on ability to use and understand verb-noun collocations. 2. Collocations that appear on the test were selected from a large corpus of written academic discourse. 3. Appropriate collocations for the collocation test have been identified. 4. Test tasks elicit performance from test-takers that reflect their collocational ability.	Domain Analysis Systematic sampling of collocations Task development

The first assumption underlying this warrant indicated that performance in written academic English tasks relies in part on ability to use and understand verb-noun collocations. Backing for this assumption came from an analysis of the domain and prior research. The target language domain was investigated through the analysis of a corpus of academic written English. Backing began with by selecting a corpus of written academic English and sampling of the collocations in the corpus. The selection process and description of the corpus were presented in chapter 3. The academic text files in the British National Corpus (BNC) were categorized as academic written English by Lee (2001). These texts were sampled as a relevant and representative sample of a variety of disciplines that students would encounter at English-medium colleges and universities.

The second assumption for this warrant states that collocations that appear on the test were selected from a large corpus of written academic discourse. This was accomplished though a systematic sampling of collocations in the target language corpus. A frequency-based approach to the identification of collocations was used as the initial step in identifying the collocations as whole units. The target corpus was processed using ConcGram 1.0 software to identify all word pairs within a span of 4 words to the right and to the left of each word in the entire 16 million-word corpus. A list was created for all pairs regardless of positional variation. A detailed description of the sampling procedure can be found in chapter 3.

The third assumption also relied on the sampling of appropriate collocations for the collocation test. Backing for this assumption came from the second half of the collocational sampling process as described in chapter 3. Once the frequency-based list of word pairs had been created, the researcher manually identified the most frequent restricted verb-noun

collocations from the list. This is called *late human intervention*, since the manual selection of collocations occurred after the list of word pairs had been obtained by the software. In this manner, high-frequency collocations that are both relevant and representative of the language in the target domain were sampled and a description of the process provides support for the second assumption that appropriate collocations for the collocations test were identified.

Finally, the fourth assumption supporting the warrant states that test tasks elicit performance from test-takers that reflect their collocational ability. Collocational ability was thus operationalized with a gap-filling task, whereby a test-taker produced an appropriate verb as a collocate for a noun as a node in the sentential context for each target collocation. Backing for this assumption came from theoretical evidence supporting the design of the task by eliciting the verb in the gap-filling task as evidence of collocational ability. Backing for this choice of task type came from previous research measuring second language vocabulary and collocational knowledge in particular, as presented in chapter 2. Further backing came from a response analysis showing that test-takers produce verbs most of the time as a response to the task.

Theoretical and empirical evidence has been found for all three assumptions that underlie the warrant for the domain description inference. The second assumption might be challenged by stakeholders in the United States, for example, because the collocations were sampled from a target domain in British English; however, the use of high-frequency collocations in academic British and academic American English tends to be similar enough to dismiss this rebuttal. The dichotomous scoring method only allowed target responses that were identified in the BNC. The polytomous scoring methods, on the other hand, awarded partial credit to responses that appeared in a corpus of contemporary American English.

Application of the polytomous scoring method would also weaken the rebuttal. Development of scoring methods was addressed in the test development chapter. Thus, sufficient evidence has been provided to support the warrant that observations of performance on the collocation test reflect the collocational ability similar to those in the target language domain of academic English and allow us to move to the next inference, evaluation.

6.1.2 Evaluation inference

The evaluation inference is based on the warrant that observations of performance on the collocation test are evaluated to provide observed scores that reflect collocational ability. There are six assumptions underlying this warrant. Evidence for these assumptions was collected through response analysis, scoring development, task trialing, descriptive statistics, and a Rash analysis. Table 6.2 shows the warrant, assumptions, and types of backing used to support the assumptions for the evaluation inference.

The first assumption underlying the warrant is that the scoring procedure captures the collocational ability learners have related to word pairs based on frequency of the whole collocation. Backing for this assumption came from an analysis of the responses. The verbs from the target collocations were used to develop the answer key and norming lists for both the dichotomous scoring method and full credit criterion for the polytomous method. These collocations were marked as a correct response in the dichotomous scale and given full credit in the polytomous scale as an indication of the verb from the collocate that was identified as a single combination in the corpus. An analysis of the responses helped to develop the answer key and identify which responses should receive partial credit with the polytomous scoring method.

Table 6.2. Warrant, assumptions, and types of backing for the evaluation inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the assumptions
Observations of performance on the collocation test are evaluated to provide observed scores that reflect collocational ability.	1. The scoring procedure captures the collocational ability learners have related to word pairs based on frequency of the whole collocation. 2. The scoring procedure accurately captures the varying degrees of strength of ability for collocations that learners have as they develop collocational ability. 3. The acceptance of misspelled words and variation in word forms results in a score that reflects collocational ability. 4. Task conditions allow test-takers to perform in a manner that exhibits what they know. 5. The statistical characteristics of items and measures are appropriate for norm-referenced decisions. 6. Items on the test fit the model predicted by a Rasch IRT analysis and measures are appropriate for distinguishing among test-takers (RQ 1)	Response analysis Scoring rubric development Task trialing Descriptive statistics IRT Rasch analysis

The second assumption required support for scoring procedure that accurately captures the varying degrees of strength of ability collocations learners have as they develop collocational ability. Backing for this second assumption was not found using the dichotomous scoring method since responses were marked only as correct/incorrect according to the target collocation. The correct/incorrect distinction did not reveal any partial

knowledge of collocational ability. The polytomous scoring method, on the other hand, allowed for partial credit which was identified though response analysis. This scale was developed as an alternative to the dichotomous scale to allow for alternate responses that are restricted by the collocability of the node in the context of a general academic domain to be identified as partial knowledge. Responses were verified in the 86-million word Corpus of Contemporary American English (COCA). Partial credit was given to responses that indicated a relationship between a collocate and a node in academic texts, thus capturing performance on a variety of collocational relationships by learners as they develop collocational ability. The partial credit awarded using the polytomous method was more sensitive to degrees of strength of collocational ability. The polytomous scoring method provided support for the assumption that scores were accurately capturing the varying degrees of strength of ability for collocations learners have as they develop collocational ability.

The third assumption was satisfied during the scoring procedure explained in chapter 4, "Methodology." Acceptable spelling deviations were added to the answer key in the learning management system (LMS) and rescored for both the dichotomous and polytomous scoring methods. In this way, both scoring methods allowed for a degree of spelling error and variation in grammatical form that resulted in scores that reflected collocational ability and not the ability to spell or produce a grammatically correct word form. This criterion in both scoring scales provided backing for the third assumption underlying the warrant.

Backing for the fourth assumption, that task conditions allowed test-takers to perform in a manner that exhibits what they know, was obtained though trialing of the gap-filling task, which focused on capturing performance of collocational ability through a discrete,

productive, context-dependent task type and response analysis. Results from the response analysis, such as the analysis in chapter 3 explaining the partial credit scoring procedure, found evidence that the majority of the responses were verbs, as expected. Since the dichotomous method only awarded credit to a single target response, the polytomous scoring method would be more appropriate for allowing test-takers to demonstrate what they know. By providing a verb in the text field on the computer, the test-taker was demonstrating a degree of collocational ability. Backing for this assumption was found due to the analysis of the responses for the discrete, productive, context-dependent task type on the test.

Backing for the fifth assumption that the statistical characteristics of items and measures were appropriate for norm-referenced decisions was found in descriptive statistics and histograms for scores using the polytomous method. Presented in Chapter 5, these data showed that the scores using the dichotomous scoring method were skewed and non-normally distributed. The distribution of scores was normally distributed using the polytomous scoring method. Consequently, the normal distribution was found to support the fifth assumption.

The sixth assumption found support by using a Rasch IRT analysis of the data. The assumption was that scores from the test items fit the model predicted by a Rasch analysis and that measures were appropriate for distinguishing among test-takers. The Rasch data showed that there were only 2 misfitting items using the dichotomous method and 8 misfitting items using the polytomous method. Further analysis revealed that half of the items for each method were overfitting the model, showing more discrimination than expected, whereas the other half were underfitting the model. The underfitting items were subject to construct irrelevant error but not enough to throw them out. In sum, only one questionable

item was found in the dichotomous data and four items in the polytomous data; thus, both methods included a high percentage of items that fit the model predicted by the Rasch analysis.

Evidence supporting the evaluation inference was presented by comparing two scoring methods. Table 5.2 shows which scoring method was favored for each assumption. Both scoring methods were found to be equivalent for three of the assumptions. Both methods awarded credit for responses related to the collocation as a whole unit, regardless of spelling or grammatical form. Most responses fit the model predicted by a Rasch analysis. Evidence for more than half the assumptions, however, favored the polytomous scoring method. This method allowed test-takers to demonstrate varying degrees of strength for collocations, allowing the test-takers to demonstrate what they know, which resulted in a normal distribution of scores appropriate for making norm-referenced decisions.

Table 6.3. Summary of support by scoring method for evaluation inference

	Supported by scoring method		
Assumption	Dichotomous	Polytomous	Both
1	X		
2		X	
3			X
4		X	
5		X	
6			X

6.1.3 Generalization inference

The next inference, generalization, based on the warrant that observed scores are estimates of expected scores which are comparable across the items on the test. This warrant

is based on the assumptions that (a) estimates of test-takers' performance can reliably distinguish among test-takers, (b) the tasks and test specifications are sufficiently detailed and consistent to produce equivalent test forms, and (c) the computer-based administration of the test is sufficiently uniform to produce consistent results. Table 6.4 shows the warrant, assumptions, and types of backing used to support the assumptions for the generalization inference.

Table 6.4. Warrant, assumptions, and types of backing for the generalization inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the assumptions
Observed scores are estimates of expected scores, which are comparable across the items on the test.	1. Estimates of test-takers' performance can reliably distinguish among test-takers (RQ 2). 2. The tasks and test specifications are sufficiently detailed and consistent to produce equivalent test forms. 3. The computer-based administration of the test is sufficiently uniform to produce consistent results.	Reliability estimates Detailed blueprints Task trialing with computer delivery

Empirical evidence for the first assumptions was sought using both scoring methods. The results from the Rasch analysis were used to partially support the first assumption. The polytomous method was found to separate the persons more reliably (.87) than the dichotomous method (.69). Additional reliability estimates also favored the polytomous method. The reliability estimate for the partial credit scale was higher ($\alpha = .89$) than for the dichotomous method ($\alpha = .83$).

Support for the second assumption was found in the detailed development of the test presented in chapter 3 and Appendix A. Equivalent test tasks and test forms should be replicable with the information provided.

Backing from assumption three came from the computer-based administration of the test, which was sufficiently uniform to produce consistent results. Consistency of the measurement was presented for assumption one, leading to consistent results. The computer-based administration delivered the test in the same way to all test-takers, allowing the same amount of time on the same size screen for each test-taker. No incidents of computer failures or trouble using the computer were reported by any test-taker.

Theoretical and empirical evidence was found to support the assumptions underlying the generalization inference. Once again, the polytomous scoring method was found to perform better than the dichotomous methods. Detailed test specifications and consistent computer-delivery of the test further supported the warrant that observed scores were estimates of expected scores which are comparable across the items on the test. Sufficient support for this warrant was found to proceed from the generalization inference to the explanation inference.

6.1.4 Explanation inference

The explanation inference is based on the warrant that expected scores can be attributed to a construct of collocational ability in academic writing, which includes knowledge of whole collocations. Five assumptions are associated with this warrant, which are supported by linguistic theory and empirical analysis. Table 6.5 shows the warrant,

assumptions, and types of backing used to support the assumptions for the explanation inference.

Table 6.5. Warrant, assumptions, and types of backing for the explanation inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the assumptions
Expected scores are attributed to a construct of collocational ability in academic writing, which includes knowledge of whole collocations.	1. Performance on the collocation test reflects test-takers' collocational ability. 2. The construct of a restricted lexical collocation has been defined as a whole collocation rather than individual constituents. 3. The more frequent a collocation, the easier the corresponding item will be for test-takers (RQ 3). 4. The scores on the Collocation Ability Test correlate as predicted to other tests of English ability related to the construct (i.e., reading and a productive vocabulary size) (RQ 4). 5. While taking the test, test-takers use metacognitive and cognitive strategies related to collocation use in academic language (RQ 5).	Response analysis Construct definition from literature Correlations with rank order of collocation Correlations with reading and vocabulary test scores Observations, interviews, and survey

The first assumption underlying this inference is that performance on the collocation test reflects test-takers' collocational ability. Backing for this assumption was found after an analysis of the responses on the test. Responses that matched the target verb in the collocation or were verified in a corpus of American academic English were found to have a

collocational relationship with the node in the sentence. High-ability test-takers were more able to produce a commutable verb with the target node reflecting collocational ability. This ability decreased with the mid-ability test-takers, and performance was even lower with low-ability test-takers. The dichotomous scoring method captured collocational ability for the target collocations; however, the polytomous method was more sensitive to measuring collocational ability at all proficiency levels.

The second underlying assumption was that the construct of a restricted lexical collocation has been defined as a whole collocation rather than individual constituents. Backing for this assumption came from the definition of collocation as a single lexical unit, as explained in chapter 2. Collocation was further defined in the literature by the limited or restricted number of word combinations that compose the collocations.

The third assumption was based on the notion that the more frequent a collocation, the easier the corresponding item would be for test-takers. Theoretically, the more frequent a word, or collocation, in a target domain, the more likely a test-taker is to know that item and potentially produce a correct collocation in context. A correlation between the rank order of the frequency of the collocations from the corpus and the mean scores on the test should have indicated a positive relationship, if this is true; however, statistical analysis comparing the item facility of each item with the rank order of the word pairs revealed little to no positive relationship between these two variables. A correlation with dichotomous scores reveled a value close to zero ($r_s = 0.2$). The correlation was only considerably higher with the polytomous scores ($r_s = .40$). A reason for these low correlations could have been due to the range of collocations. Based on the idea of frequency as an important factor in designing a test of collocational ability, high-frequency collocations were identified and selected as target

collocations. As a result, the collocations on the test were fairly frequent, whereas infrequent collocations were not selected. As a consequence, the range of collocations as represented by rank order was severely restricted. The range of scores, while not represented by the rank order scale used for this statistical analysis, may have played a role in the facility of the item on the test. As a result, backing for this assumption is minimal and only associated with the polytomous method.

Backing for assumption four began in chapter 2 with a discussion of the theoretical relationship among reading, vocabulary, and collocational ability. Empirical evidence was presented in chapter 5 showing correlation estimates among collocational ability, reading, and productive vocabulary. Partial backing was found for the assumption that scores on the collocation ability test correlated as predicted to other tests of English ability. Similar results were found for each scoring method using correlations corrected for attenuation to compare true scores of the theoretical constructs. A positive correlation was found between dichotomous scores on the Collocational Ability Test and scores on the reading test ($r_s = .70$) as well as the polytomous scores on the Collocational Ability Test and the reading test ($r_s = .62$). Total scores on the Collocational Ability Test also correlated positively with total scores on the vocabulary size test for both the dichotomous scale ($r_s = .74$) and the polytomous scale ($r_s = .72$).

Backing for the fifth assumption that while taking the test, test-takers use metacognitive and cognitive strategies related to collocation use in academic language, was found through the responses on screen recordings during test administration, the Test Response Survey, and post-test interviews. An analysis of video files of a sample of test-takers indicated that high-ability test-takers adjusted their responses twice as often (M =

4.75) as mid- (M = 2.5) and low-ability (M = 2.5) test-takers. Changes to responses were seen as evidence that the test-takers were engaged in the language and using metacognitive and cognitive strategies to identify an appropriate response. Further analysis looked at (a) the degree to which test-takers were thinking about academic English as they took the test and (b) the comparison with English on the test and in university textbooks. The results of a chi-squared test showed that responses to the test reflection survey indicated that the groups did not differ in while thinking about academic English while taking the test. A statistical difference was found in their perception regarding a comparison of English on the test and academic English in university textbooks. The high-ability group had the largest percentage of test-takers who were able to compare the text on the test with English in university textbooks. The mid-ability group was next, followed by the low-ability group. Overall, for the test-takers who were able to make a comparison, about half of the test-takers in all three groups reported that the texts were different, and half reported similarities. Test-takers in the high- and mid-ability groups pointed to differences related to academic disciplines, sentence structure and vocabulary. Responses in the low-ability group were descriptions of discourse level characteristics such as those taught in an English language course without actually having much experience reading or writing academic English. Overall, the test-takers in the high-ability group were more likely to see a similarity or difference between the academic English on the test and in university textbooks. They were also more articulate with specific details when commenting on these comparisons. The mid- and low-ability test-takers were less able to make a comparison.

Post-test interviews with an additional sample were used to explore the differences among the groups. Comments from the post-test semi-structured interviews for a sample of

test-takers revealed that performance may be attributed to awareness of test purpose and explicit learning of phrases in English. Test-takers with higher performance on the test reported studying word pairs as they learned English. There was evidence of blending collocations. This occurred when a test-taker produced a verb for a collocation that was not appropriate. The verb was familiar because the test-taker had knowledge of another similar collocation that did use that verb. As a contrast, no interviewee with low performance acknowledged memorizing lists or word pairs or studying collocations for English language instruction or having much knowledge of the collocations on the test. Furthermore, high-ability test-takers appeared to use more cognitive and metacognitive strategies when taking the test; however, awareness of the characteristics of academic English was weak at all levels.

Evidence that was found backing the assumptions supporting the warrant for the explanation inference varied. Theoretical evidence supported the construct of collocational ability, which was used to support the first two assumptions. The polytomous method was superior to the dichotomous method when compared with rank order of collocations but not with other correlations of English ability. Moderate correlations with scores on reading and vocabulary tests were found as predicted. Evidence of cognitive and metacognitive strategies was found in the results from the qualitative data collection, and high-ability test-takers were more aware of academic language and appeared to evaluate the language on the test and adjust their responses accordingly. On the other hand, little difference was found between the scoring methods and the relationship with reading and productive vocabulary size. Overall, the qualitative data supported the results from the quantitative data indicating that high-

ability test-takers were more familiar with collocations in academic written English and were better at producing a missing verb in context and used more strategies to do so.

6.1.5 Extrapolation inference

The next inference in the argument is the extrapolation inference. This inference is based on the warrant that the construct of collocational ability as assessed by the Collocation Ability Test accounts for the academic collocational language performance in academic discourse in English-medium colleges and universities. The assumptions are that (a) the collocations appearing on the test reflect those that the test-takers will find in an academic context, (b) scores on the collocation test distinguish among proficiency groups with and without experience and knowledge of academic language, and (c) scores on the collocation test have a positive relationship with scores on other measures of academic vocabulary. Table 6.6 shows the warrant, assumptions, and types of backing used to support the assumptions for the extrapolation inference.

Backing for the first assumption came from the selection and description of target domain as represented by the target language corpus and the American English corpus that was used to verify results for partial credit scoring described in chapter 3. Collocations were identified that were relevant and representative of the collocations that are used in written academic language that is representative of the language used in academic disciplines. The language is relevant to academic discourse rather than conversation language that might take place in a laboratory or conference. This is the language that the test-takers will encounter and use in academic study at an English-medium college or university.

Table 6.6. Warrant, assumptions, and types of backing for the extrapolation inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the inferences
The construct of collocational ability as assessed by the Collocation Ability Test accounts for the academic collocational language performance in academic discourse in English-medium colleges and universities.	1. The collocations appearing on the test reflect those that the test-takers will find in an academic context. 2. Scores on the collocation test distinguish among proficiency groups with and without experience and knowledge of academic language (RQ 6) . 3. Scores on the collocation test have a positive relationship with scores on other measures of academic vocabulary (RQ 7) .	Domain analysis Analysis of variance Correlation with academic vocabulary sub-test

The second assumption was supported empirically by an ANOVA which showed that groups were statistically different from one another. The groups in the study, described in chapter 3, were selected to represent a variety of proficiency levels based on placement and degree of matriculation at an English-medium college or university. Results using a one-way independent ANOVA indicated that the three groups were significantly different for both the dichotomous method $F(2, 203) = 91.93$, $p < .0001$ and the polytomous method $F(2,203) = 64.50$, $p < 0.0001$. The high-ability group of test-takers, who were already working and studying at a graduate level, performed significantly better than the other two groups. The next group consisted of students who had been admitted to the university yet still were enrolled in additional instruction in academic English language, followed by the low-ability test-takers who have had the least exposure and experience with academic language at an English-medium institution. The significant differences among groups provided support to

the assumption that scores on the collocational test can distinguish among proficiency groups.

Finally, the assumption that scores on the collocation test have a positive relationship with scores on other measures of academic vocabulary was investigated with evidence that was found in correlations between scores on the Collocational Ability Test and the sub-test of the vocabulary test for both scoring methods. After correcting for attenuation, the correlation coefficient was higher between the dichotomous scores on the Collocational Ability Test and scores on the sub-test of high-frequency vocabulary ($r_s = .70$) than between the dichotomous scores and the scores on the sub-test of academic English ($r_s = .68$).

The results were similar using the polytomous data. For the polytomous scoring method, the correlation with the vocabulary sub-test developed using high-frequency vocabulary was greater ($r_s = .73$) than the sub-test of high-frequency vocabulary ($r_s = .70$). Both scoring methods resulted in higher correlations with measures of high-frequency vocabulary than with measures of academic vocabulary; however, the correlations were fairly high for both measures. Rather than interpret the results as a greater relationship with high-frequency vocabulary, one therefore could say that the collocations on the Collocational Ability Test have a positive relationship with both academic and high-frequency vocabulary.

Evidence was found to support the assumptions supporting the extrapolation inference. A relationship with a measure of academic vocabulary was not found to be greater than a relationship with high-frequency for either scoring method but was fairly high for both measures. Finally, both methods were able to distinguish among groups with differing placement and proficiency levels at the university.

6.1.6 Utilization inference

The utilization inference in the validity argument for the Collocational Ability Test rests on the warrant that performance on the test contributes to making appropriate decisions about matriculation and placement in English-medium colleges and universities. This warrant would find support from the assumptions that (a) score-based interpretations provide enough information to contribute to the decision making process, and (b) test scores contribute to and facilitate student placement in English language courses at English-medium colleges and universities. Table 6.7 shows the warrant, assumptions, and types of backing used to support the assumptions for the utilization inference. Types of backing are listed in parentheses since evidence has not yet been obtained.

Table 6.7. Warrant, assumptions, and types of backing or the utilization inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the assumptions
Performance on the test contributes to making appropriate decisions about matriculation and placement in English-medium colleges and universities.	1. Score-based interpretations provide enough information to contribute to the decision making process. 2. Test scores contribute to and facilitate student placement in English language courses at English-medium colleges and universities.	(Successful score interpretation) (Correlations with course grades)

An exploratory analysis indicated that a number of test-takers in each group would have been placed differently if placement decisions were based solely on the scores from the Collocational Ability Test. Once the test is operational, additional backing for these two

assumptions can be sought. The first assumption can be supported by the interpretation of the scores and their corresponding interpretation with the English language courses offered by the college or university. Backing for the second assumption could be found during the decision making process if the additional information provided by the scores contribute to clearer cut-scores for placement.

6.1.7 Impact intention inference

The impact intention inference is based on the warrant that test score interpretation and use is beneficial for all test users and stakeholders. Table 6.8 shows the warrant, assumptions, and types of backing used to support the assumptions for the impact intention inference. Types of backing are listed in parentheses since evidence has not yet been obtained.

Table 6.8. Warrant, assumptions, and types of backing for the impact intention inference

Warrant licensing the inference	Assumptions underlying the warrant	Types of backing used to support the assumptions
Test score interpretation and use is beneficial for all test users and stakeholders.	1. The test construct raises awareness about the importance of collocations in academic English. 2. Instructors are aware of the potential benefits of alternate scoring methods for constructed response tasks.	(Washback studies) (Stakeholder surveys)

The first assumption supporting this warrant states that the test construct raises awareness about the importance of collocations in academic English. Backing for this

assumption can come from washback studies looking at course syllabi and language learning materials used in the courses. The second assumption is that instructors are aware of the potential benefits of alternate scoring methods for constructed response tasks. Backing can be provided by responses on a survey conducted with English-language instructors regarding the advantages and disadvantages of the two scoring methods.

6.2 Summary of validity argument

This study has provided evidence to support five of seven inferences in the interpretive argument as the basis for the development of a validity argument for the meaning of a score on an ESL collocational ability test based on corpus-driven design. Theoretical and empirical evidence has supported the assumptions related to the domain definition, evaluation, generalization, explanation, and extrapolation inferences. Support for the utilization and impact intention inferences can be collected after the test is operational. Evidentiary support for the assumptions underlying each inference provides justification for score interpretation that is needed for the collocational ability test so that a score can be used to contribute to the decision to allow a test-taker to participate in English-medium instruction or place the test-taker in an appropriate English language course.

6.3 Limitations and implications

The limitations in this study are related to sample size of participants for the qualitative data collection and to the evidence needed to complete the support for the interpretive argument.

The first limitation is the small sample size for the qualitative data analysis. Ten participants were included in the screen capturing data and only six participants took part in the post-test interviews. The data obtained from these two sources were enlightening and helped support the findings from the quantitative data; thus, it would have been beneficial to include more participants in the qualitative data collection.

A second limitation to this study was the inability to find support for all of the inferences in the interpretive argument. Warrants and assumptions were developed for the utilization and impact intention inferences, yet it is difficult to collect empirical evidence to support the assumptions underlying these inferences until the test is actually used to make placement decisions. This limitation directly affects the purpose of the argument-based approach to validation, which evaluates the interpretations of score meaning and score use. Since the scores on the Collocational Ability Test are intended to be used as part of the decision making process to place students into appropriate courses at an English-medium college or university, it is necessary to support the assumptions underlying the warrants for the utilization and impact intention inferences.

The findings from this study have implications for stakeholders: language learners, materials developers, language classroom instructors, and test developers. In each situation, technology can be useful in assisting the stakeholders with the benefits from this study.

The first implication is an awareness of restricted verb-noun collocations and phraseology in written academic English. Language learners can benefit by noticing the words that make up a collocation as a whole unit in a particular context. Attention should also be paid to the distribution of collocations by register such as written academic English. Classroom instructors can assist their students in becoming aware of academic collocations

through lesson plans and appropriate materials selection. Large language corpora are available and can be used to identify or verify language features in context. These corpora can be used by the classroom instructor to plan a lesson or by the student to learn or verify word combinations in context. Materials developers have already begun to include activities with collocations in textbooks. The connection between what is being taught and what is being tested will become stronger with an awareness of collocations in written academic English and other contexts.

A second implication is a discussion of the construct of collocational ability. This has an impact on how and what students learn, instructors teach, and materials and test developers include in materials and assessment instruments. The first part of the discussion questions whether collocations are co-selected individual words or a single lexical unit. This is followed by an awareness of procedures for sampling collocations and factors in addition to frequency that might affect collocational ability. The definition of a collocation has an impact on how collocations are sampled in a language domain. The sampling procedure has implications for how collocations are learned, taught, and tested.

A third implication is a focus on measuring partial collocational knowledge with gap-filling tasks. The measurement of partial knowledge in this study was accomplished with a polytomous scoring method. The dichotomous scoring method only awarded credit for responses that matched the target collocations. With the polytomous method, test-takers were given credit for knowledge of lexical relationships other than the target collocation. This is a more sensitive measure of collocational ability.

Finally, this study illustrated the benefits of using an argument-based approach to validation. This approach was useful for evaluating the Collocational Ability Test and

comparing scoring methods and has an effect on how tests are or should be evaluated. An evaluation of the collocation test using a different approach to validation might have favored the dichotomous scoring method. Consequences of this approach, a less sensitive measure of collocational ability and inability to make norm-referenced decisions because of the non-normal distribution of scores, might have been dismissed.

6.4 Suggestions for future research

My hope is that this study will raise awareness about the potential for collocational ability to be used to strengthen placement decisions and inform classroom instruction and materials development. Replication studies involving a sample of participants from other populations and with other characteristics (i.e., a variety of L1s) may result in tenable alternatives for evidence in a validity argument for a test of collocational ability.

Other future research can further the systematic sampling procedure through attention to the selection of corpora and the method for identifying target collocations. This study has found a relatively weak link between item facility and frequency by rank order due to the restricted range of the target collocations. Research examining the role that frequency plays in alternative formulas effecting item difficulty would be beneficial.

Another study can focus on the concept of lexical collocations as a single unit or single lexeme (Schmitt, personal communication). Collocations might be learned or stored in the mind as single units or as relationships between lexical items. These connections might be different when learning a second language than when learning a first language. If we can establish how learners acquire collocations, our measurement of collocational ability might

improve. The findings would have direct implications on how the framework and construct of collocational ability is defined and how collocational ability is assessed and acquired.

Finally, continued development of argument-based validation of tests measuring collocational ability beginning with the purpose of the test (i.e., test uses and intended beneficial consequences) will unite our research efforts making the interpretations of our findings more meaningful and comparable.

6.5 Conclusion

This chapter began with an overview of research, beginning with the development of the collocation test, and continued with a description of the materials and methods used. This study has developed a validity argument for a computer-delivered ESL test of collocational ability. The validity argument was an evaluation of the theoretical analysis and empirical data that were found to support the warrants in the interpretive argument for the Collocational Ability Test. The chapter concluded with limitations and implications of the study and suggestions for future research.

References

Aghbar, A. A. (1990). *Fixed expressions in written texts: Implications for assessing writing sophistication.* Paper presented at the Annual Meeting of the Conference on College Composition and Communication, Detroit, MI, March 17-19.

Aghbar, A. A., & Tang, H. (1991). *Partial credit scoring of cloze-type items.* Paper presented at the 1991 Language Testing Research Colloquium, Educational Testing Service, N.J.

Aisenstadt, E. (1979). Collocability restrictions in dictionaries. *ILT: Review of applied linguistics, 45*(6), 71-74.

Aisenstadt, E. (1981). Restricted collocations in English lexicology and lexicography. *ILT: Review of Applied Linguistics, 53*, 53-61.

Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System 38*, 391-401.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Aston, G. (2001). (Ed.) *Learning with corpora.* Houston, TX: Athelstan

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, UK: Cambridge University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford, UK: Oxford University Press.

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal, 47*, 56-63.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101-114

Ball, F. (2001). Using corpora in language testing. *Cambridge ESOL Research Notes, 6*, 6-8.

Barker, F. (2004). Using corpora in language testing. *Modern English Teacher, 13*(2), 63-67.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper.* TOEFL Monograph Series, Number 19. Princeton, NJ: Educational Testing Service.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English*. Philadelphia, PA: Benjamins.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: A Polish/German empirical study. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 85-93). London, England: Macmillan.

Blum-Kulka, S., & Levinson, E. (1983). Universals of lexical simplification. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 119-139). London, England: Longman.

Bond, T. G., & Bond. C. M. (2007) *Appling the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In J. D. Brown & T. Hudson (Eds.), *A focus on language test development: Expanding the language*

proficiency construct across a variety of tests (Technical Report #21) (pp. 113-142). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 171-174.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York, NY: McGraw-Hill.

Carr, N. (2011). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

Catch. (2010). In *Oxford dictionaries online*. Retrieved from http://oxforddictionaries.com /definition/catch

Channell, J. (1981). Applying semantic theory to vocabulary teaching. *English Language Teaching Journal, 35*(2), 115-122.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.) *Second language acquisition and language testing interfaces*. Cambridge, UK: Cambridge University Press.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. E. Enright & J. Jamieson (Eds.) *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-350). New York, NY : Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language.* London: Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chen, K-J., Huang, C-R., Chang, L-P, & Hsu, H-L. (1996). Sinica corpus: Design methodology for balanced corpora. In. B.-S. Park and J.B. Kim (Eds). Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp.167-176.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411–433.

Cheng, W., Greaves, C., Sinclair, J. McH., and Warren, M. (2009). Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics, 30*(2): 236-252.

Clear, J. (1993). From Firthian principles: Computational tools for the study of collocation. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 271-292). Philadelphia, PA: Benjamins.

Collins COBUILD English collocations [CD-ROM].(1995). Birmingham, England: HarperCollins.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 149-161). Philadelphia, PA: Benjamins.

Creswell, J. W., & Plano Clark, V. L., (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Crowther, J. (Ed.). (1995). *Oxford advanced learner's dictionary of current English* (5th ed.). Oxford, England: Oxford University Press.

Davies, M. (2008) *The corpus of contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Douglas D. (2000). *Assessing languages for specific purposes.* Cambridge, UK: Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. Oxford, UK: Oxford University Press.

Durrant, P. (2008). *High frequency collocations and second language learning.* PhD thesis, University of Nottingham.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28,* 157-169.

Durrant, P. & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of lexical priming. *Corpus Linguistics and Linguistic Theory 6*(2), 125-155.

Ellis, N., Frey, E. & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody. In U. Romer and R. Schulze (Eds.) *Exploring the lexis-grammar interface.* (pp. 89-114). Amsterdam, John Benjamins Publishing Company.

Enright, M, K., Grabe, W., Koda, K., Mosenthl, P., Mulcahy-Ernt, P., & Schedl, M. (2000). TOEFL 2000 reading framework: A working paper (TOEFL Monograph no. 17). Princeton, NJ: Educational Testing Service.

Eyckmans, J. (2009). Toward an assessment of learners' perceptive and productive syntagmatic knowledge. In A. Bafield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretation* (pp. 129-154). New York, NY: Palgrave MacMillan.

Farghal, M., & Obiedat, H. (1995). Collocations: A neglected variable in EFL. *International Journal of Applied Linguistics, 28*(4), 313-331.

Fletcher, W. H. (Ed.) (2003). *Phrases in English*. Retrieved from http://phrasesinenglish.org/

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Oxford, UK: Routledge.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey, *Applied Linguistics, 28*(2), 241-265.

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. San Francisco, CA: International Scholars.

Grabe, W. (2009) *Reading in a second language: Moving from theory to practice.* Cambridge, UK: Cambridge University Press.

Greaves, C. (2009). ConcGram 1.0: A phraseological search engine [Computer software].

Greenbaum, S. (1988). Some verb-intensifier collocations in American and British English. In S. Greenbaum (Ed.), *Good English and the grammarian.* (pp. 113-124). London, England: Longman.

Gyllstad, H. (2005). *Words that go together well: Developing test formats for measuring learner knowledge of English collocations*.

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocational knowledge: COLLEX and COLLMATCH. In A. Bafield & H. Gyllstad (Eds.) *Researching*

collocations in another language: Multiple interpretations (pp. 153-170). New York, NY: Palgrave MacMillan.

Hambelton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Nijoff.

Hellekjær, G. O. (2009). Academic English reading proficiency at the university level: A Norwegian case study. *Reading in a Foreign Language, 21*(2), 198-222.

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies, 16*(3), 380-384.

Howarth, P. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making.* Tübingen, Germany: Niemeyer.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics 19*(1). 24-44.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.

Iwashita, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., & Tono, Y. (2003). *JACET 8000: JACET list of 8000 basic words*. Tokyo, Japan: JACET.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 19*(2), 219-239.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed.), (pp. 17-64). Westport, CT: Greenwood.

Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17*(1), 81-92.

Koyo, T. (2003). A study of collocation in English and Japanese noun-verb combinations. *Intercultural communication studies, 12*(1), 125-147.

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with readers general academic ability? *Journal of Research in Reading, 15*(2), 95-103.

Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, *5*(3), 37-72.

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

Martinez, R. (2010, April). Evidence of lack of processing of multiword lexical items in reading tests. Paper presented at the Language Testing Research Colloquium (LTRC). Cambridge, England.

Meara, P. (1996). The vocabulary knowledge framework. Available: http://www.lognostics.co.uk/vlibrary/meara1996c.pdf

Meara, P. (1997). Towards a new approach to modeling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.) *Vocabulary: description, acquisition and pedagogy.* (pp. 109-121). Cambridge: Cambridge University Press.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 47-58). Cambridge, UK: Cambridge University Press.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-62.

Mochizuki, M. (2002). Exploring two aspects of vocabulary knowledge: Paradigmatic and collocational. *Annual Review of English Language Education in Japan, 13,* 121-129.

Moreno Jaén, M. (2007). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International Journal of English Studies, 7*(2), 127-147.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24,* 223–242.

Oakey, D. (2009). Conceptual and methodological approaches to word combinations. In S. Hunston & D. Oakey (Eds.), *Introducing applied linguistics concepts and skills* (pp. 29-58). London, England: Routledge.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191-226). New York, NY: Longman.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Revier, R. L. (2009). Evaluating a new test of whole English collocations. In J. Gyllstad & Barfield, A. (Eds.) *Researching collocations in another language: Multiple interpretations* (pp. 125-138). New York, NY: Palgrave Macmillan.

Rudzka, B, Channell, J., Putseys, Y, & Ostyn, P. (1981). *The words you need*. London, England: Macmillan.

Schmitt, N. (1998, June). Measuring collocational knowledge: Key issues and an experimental assessment procedure. *ILT: Review of Applied Linguistics, 119-120,* 27-47.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Grandage, S. & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.) Formulaic Sequences: Acquisition, processing and use. (pp. 127- 151). Amsterdam: John Benjamins Publishing Company.

Seidl, J. & McMordie, W. (1978). *English idioms and how to use them*. Oxford, UK: Oxford University Press.

Shillaw, J. (2009). Commentary on part III: Developing and validating tests of L2 collocation knowledge. In A. Bafield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 171-177). New York, NY: Palgrave Macmillan.

Sinclair, J. (Ed.). (1995). *Collins COBUILD English dictionary* (2nd ed.). London, England: HarperCollins.

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* Oxford, UK: Oxford University Press.

Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J. R. Firth* (pp. 410-430). New York, NY: Longman.

Sinclair, J. M. (2005). Developing linguistic corpora: A guide to good practice corpus and text-basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk /creating/guides/linguistic-corpora/index.htm

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429-458.

Steiger, J. H. (1980). Tests for comparing elements of a correlational matrix. *Psychological Bulletin, 87*(2), 245-251.

Summers, D. (Ed.). (1995). *Longman dictionary of contemporary English* (3rd. ed.). London, England: Longman.

Vermeer, A. (2001). Breadth and depth of vocabulary relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217-234.

Voss, E. (2008). *Measuring knowledge of lexical collocations: Comparing item types on web-based tests*, Paper presented at MWALT 2008, Iowa City, Iowa, September 20, 2008.

Webb, S., & Kagimoto E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly, 43*(1), 55-77.

Wolter, B. & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics,* 1-21.

Appendix A: Summary of task characteristics and variables

Task characteristics	Task variables	Definition of variable
Situation	Participants	Mandarin-speaking learners of English who are preparing for study at an English-medium college or university
	Content	Published and unpublished texts consisting of six academic disciplines, technical engineering, social science, politics & law, natural science, medicine, and humanities & arts
	Setting	Academic
	Purpose	Primarily representational: Conveying meaning about academic topics and content
	Register	Formal written academic English
Text collocations material	Language Features	High-frequency restricted verb-noun
	Pragmatic Features	Ideational functions: to express or exchange information about ideas and knowledge
	Discourse Features	Knowledge of genre, register, and collocation
Test rubric	Questions/ Directives	Fill in the gap with a verb that completes the sentence with an appropriate meaning for the academic context
	Response Formats	Sentential gap-filling: Produce a missing verb to complete the appropriate collocation in a sentence for academic English.
	Rules for Scoring	Target responses (collocations with the highest raw frequency in the corpus) will receive full credit (2 pts). Optional: Partial credit (1 pt) will be awarded to commutable verbs with a raw frequency of 5 instances or higher verified in COCA.

Appendix B: Test items

1. A curriculum-based report is likely to _____ a greater chance of success because it shows benefits to pupils, the teachers and the school librarian i.e., the whole school. [key: have] – 28 words
2. A distinction can be _____ between planned and unplanned decentralization. [key; made] – 10 words
3. A number of steps can be _____ to speed up processing, each depending on some different method of handling the search. [key: taken] – 21 words
5. All patients were _____ with omeprazole 40 mg/day for a median duration of 12 weeks. [key: treated] – 15 words
6. Although it might be very easy to steal from a friend or colleague and not get caught most people would feel this to be "wrong", yet such feelings may not _____ such a strong influence over decisions as to whether to steal from a larger and less personal victim. [key: exert] – 49 words
7. Although kings _____ sporadic efforts to reform, to Purge and reorganize, again ultimately only the nineteenth century brought the abundance which permitted a system to organize the State and curb some of the worst excesses. [key: made] – 35 words
10. Claudius _____ two other arrangements which seem at first sight to be highly anomalous. [key: made] – 14 words
11. Coastal configuration (e.g., funnel-shaped estuaries), deltaic sedimentation and tidal movement may also _____ a part in creating local differences in sea-level rises. [key: play] – 22 words
12. Data will be _____ by means of questionnaires to mother and fathers and interviews with mothers and siblings. [key: collected] – 18 words
15. For the most part the school _____ little control over these types of evaluation so we will consider the approaches in outline, concentrating on the issues raised. [key: has] – 27 words
17. Further information was _____ from performance indicators from the Department of Health. [key: obtained] – 12 words
18. I do not believe that linguistics has any contribution to _____ to the teaching of English or the standard European languages. [key: make] – 21 words
19. I have therefore _____ the conclusion that judicial review does not lie to impeach the decisions of a visitor taken within his jurisdiction (in the narrow sense) on questions of either fact or law. [key: reached] – 34 words
21. If the person is a juvenile or is mentally disordered or mentally handicapped the notice shall be _____ to the appropriate adult. [key: given] – 25 words
22. If this results in a tie, an arbitrary choice may be _____. [key: made] – 12 words

23. Impressive results were _____, with only 5 out of 243 noun phrase brackets being omitted. [key: obtained] – 15 words
24. In all these cases, the Attorney-General is not bound to _____ legal action, even if the law has clearly been broken. [key: take] – 21 words
25. In order to achieve this balance care must be _____ in the selection of sources from which the corpus is created and how much of each source is used in the corpus. [key: taken] – 32 words
26. It is essential for the powerless and the poor to _____ access to as large a range of legal services and skills as those at the disposal of the authorities. [key: have] – 30 words
27. It is important to _____ a clear understanding of the impact of legal aid on tribunal representation. [key: have] – 17 words
28. Little attention has been _____ to the relationship between mental and social class for the older age groups. [key: given] – 18 words
29. Morbidity is a broad term _____ to describe non-mortal patterns of ill health within the population. [key: used] – 16 words
30. New problems may _____ the form of crises, such as sudden and/or final loss of ability to walk, or the development of pressure sores. [key: take] – 24 words
31. Only when this has been done, can any serious attempt be _____ to analyze any crucial stratified groups from civil or military site. [key: made] – 23 words
32. Parents use a variety of techniques, from reasoning and explaining to threatening and slapping, when a child is having a tantrum, all of which _____ little or no effect. [key: have] – 29 words
33. Particular attention will be _____ to differences in political culture and the role of linkages between local and regional interest groups in political developments. [key: paid] – 24 words
35. Secondly, as already mentioned, some consideration must be _____ to the status of the sites being compared in the pattern. [key: given] – 20 words
38. The motivations - both many and complex - behind Japan's desire to exert her influence in Asia have _____ considerable attention from historians. [key: received] - 23
39. The proportion of the different grains present is more difficult to ascertain and is done _____ two main techniques, point counting and visual estimates. [key: using] – 24 words
41. The time constraints for decisions to be _____ are tight, as operators are required to give only six weeks' notice of their intention to withdraw or change a service. [key: made] – 29 words
42. There are several ways of analyzing the services _____ by local government. [key: provided] – 12 words

44. This contract is _____ between ourselves and yourselves as principals, we alone being liable to you for its performance. [key: made] – 19 words
45. This means that statements can be _____ describing typical differences in wealth between the country groups without needing to mention the differences in spread in the same breath. [key: made] – 28 words
48. When, later in the poem, reference is _____ to cultural and spatio-temporal elements, such qualification is missing. [key: made] – 17 words
49. Would it _____ any difference if the second extract was rewritten with referring expressions instead of repetition? [key: make] – 17 words

Appendix C: Target verb-noun word pairs sampled from

BNC academic sub-corpus with most frequent collocate form

1. Have an effect - have
2. Make a decision - made
3. Provide evidence – provide
4. Collect data – collected
5. Attract attention – attracted
6. Receive attention - received
7. Adopt an approach- adopt
8. Pay attention – paid
9. Reach a conclusion – reached
10. Play a part – play
11. Spend time – spent
12. Use technique – using
13. Obtain results - obtained
14. Provide support - provided
15. Play a role – play
16. Draw attention - draw
17. Make contribution - make
18. Provide service - provided
19. Have opportunity - have
20. Make sense - make
21. Have control - has
22. Make distinction - made
23. Have authority - has
24. Make attempt - made
25. Treat patient - treated
26. Take decision - taken
27. Have advantage - has
28. Use term - used
29. Have access - have
30. Take step - taken
31. Give attention - given
32. Have consequence - have
33. Take care - taken
34. Take form - take
35. Take action - take
36. Give notice - given
37. Use information - used
38. Make reference - made
39. Obtain information - obtained
40. Cause damage - caused
41. Make effort - made
42. Have understanding - have
43. Make contract - made
44. Have chance - have
45. Give consideration - given
46. Make difference - make
47. Make statement - made
48. Make arrangements - made
49. Make choice - made
50. Exert influence - exert

160	the diet of scholasticism he was fed there. His	time	was often	spent, he said, looking at maps in the
161	in a disposition essentially formless so much	time	should be	spent by the jurists on questions of
162	carers who were managing to retain full-	time	jobs often	spent a considerable amount of money
163	he hires or buys the goods. However, a little	time	should be	spent considering the position of the
164	for his difficult behaviour. The only	time	the mother	spent with the boy was during
165	probability density function for the amount of	time	the signal	spent away from the electrical
166	entitled to costs in respect of their own	time	and effort	spent preparing and presenting the
167	About one third of your daily language learning	time	should be	spent in this activity. It may take
168	individual, including financial expenditure,	time	and effort	spent. Therefore, the availability,
169	to make housework pleasant by lightening the	time	and effort	spent doing it (1934 p 32). The
170	only allow up to £7.50 per hour for the	time	reasonably	spent in preparation. Costs
171	worse because, in many schools, most of the RE	time	available is	spent on content only marginally
172	considerable advancement in agriculture. Their	time	is constantly	spent in tilling the soil,
173	SSE but many raised doubts about whether such	time	would be well	spent. As one teacher said,
174	are preoccupied with the increasing amount of	time	which must be	spent in the office with paper
175	And who will bear the cost? Is the money and	time	which will be	spent on preparation, classroom
176	political. From about 1864 on, much of Marx's	time	and energy was	spent in connection with the
177	details which fascinated me. So much	time	and thought was	spent in working these out and
178	school age and a significant amount of their	time	will therefore be	spent at school. Teachers
179	experience of Kadiri Celebi, the Mufti in the	time	of Suleyman, who	spent nine years in great
180	involvement of parents, a head's disposable	time	could be better	spent in grappling at his or
181	community that the later years of life were a	time	of &quotall passion	spent&quot &mdash that sex
182	on the theme &bquoPacking&equo, children	spent	time	discussing occasions when they packed and
183	harder for it. The owner will have	spent	time	finding out about his territory; where the best
184	colleges and environmental groups have	spent	time	in remote islands, producing for their own use
185	Secretary Humphrey, for instance,	spent	time	with Macmillan trying to find ways to protect
186	week before my next visit. 11.35:	Spent	time	with J. Talked about how his day was going,
187	of practice activity before 1990 and	spent	time	equipping general practitioners instead of
188	President Mitterrand's Government, had	spent	time	with Guevara in the jungle, and had been
189	But those least likely to have	spent	time	in a hospital or hospice or to die in one were
190	of Bishop Hermann of Sherborne, probably	spent	time	in the community there, and was writing at
191	Independence; many of the new leaders had	spent	time	in gaols, and devoted some attention to the
192	fewer of all those aged 85 or more had	spent	time	in hospital &mdash 64 per cent against 80 per
193	Clarkson and in 1822 William Allen all	spent	time	at the congresses meeting political leaders and
194	position. They found that when women	spent	time	out of paid work bringing up their children they
195	with some of those bishops who had	spent	time	in the island monastery. Taken together, this
196	directly from school, some will have	spent	a time	in business and will be wishing to improve their
197	given very little account of how he	spent	his time,	but he has said he spent the majority of that
198	few, one in ten, of those who had	spent	any time	in residential homes were thought to have been
199	once more at Sorgues, while Picasso	spent	his time	between Sorgues and Avignon, where he worked in
200	showed that married people less often	spent	any time	in residential homes than single or previously
201	near Indian settlements. They	spent	the time	there swimming and walking, and when the
202	in the local community. Instead, he	spent	his time	taking apart clocks and listening to Western
203	was a quarter for those who had not	spent	any time	in a residential home. Residential homes,
204	He stayed in Paris for a week, and	spent	the time	looking at paintings and collecting souvenirs

Appendix E: COCA results and test-taker responses for "make ~ distinction"

Collocate	Frequency	Test-taker Responses	Credit		Collocate	Frequency	Test-taker Responses	Credit
IS	284	a link	0		INVOLVES	7	gap	0
MAKE	189	acsadnt	0		SHOW	7	generated	0
MAKES	90	appeared	0		USE	7	get	0
MADE	78	approved	0		ACCEPTED	6	great	0
BE	69	aroused	0		APPLY	6	happen	0
WAS	63	aviod	0		APPRECIATE	6	happened	0
HAVE	56	brought	0		CONSIDER	6	have	1
DRAW	53	Calculated	0		CREATED	6	identified	0
HAS	53	change	0		FOUND	6	important	0
MAKING	45	changed	0		SEEM	6	in	0
DRAWS	35	choose	0		ACHIEVE	5	indentied	0
DREW	29	choosen	0		ARGUES	5	initiate	0
CAN	27	chosen	0		AWARDED	5	known	0
ARE	26	classified	0		BEING	5	long	0
DRAWING	23	classify	0		COULD	5	made	2
BLUR	22	clear	0		CREATES	5	make	2
BLURS	21	compare	0		DISCUSSED	5	maken	2
BASED	20	compared	0		DRAWN	5	making	2
DOES	19	coneccet	0		ELIMINATE	5	marked	0
WOULD	17	confuse	0		ESTABLISH	5	measured	0
BLURRED	16	consider	1		IGNORE	5	notice	0
'S	16	created	1		IMPLIES	5	noticed	0
HAD	15	decide	0		INDICATING	5	observed	0
CLARIFY	14	decided	0		NOTE	5	obtained	0
SEE	14	define	0		POINT	5	obvious	0
DO	13	defined	0		POINTS	5	occur	0
MAINTAIN	13	definited	0		PROPOSED	5	occured	0
MAINTAINED	13	depended	0		PROVIDES	5	occurs	0
WILL	13	deperated	0		REFLECT	5	outbreak	0
MIGHT	12	describe	0		RELIES	5	placed	0
RECOGNIZE	12	designed	0		REQUIRE	5	provided	1
SUGGESTS	12	detected	0		RETAINS	5	put	0
ARGUED	11	determined	0		REVEALED	5	raised	0
MUST	11	developed	0		SERVED	5	realized	0
NOTED	11	devided	0		STRESS	5	recognation	0
SUGGEST	11	differ	0		TAKE	5	recognized	0

BEEN	10	differed	0		USED	5	reconozed	0
BLURRING	10	difference	0				recorganized	0
LIES	10	different	0				Return	0
MAINTAINING	10	differentiate	0				Searched	0
RECOGNIZED	10	differentiated	0				See	1
SHOULD	10	distincted	0				Seen	1
UNDERSTAND	10	distinguish	0				selected	0
COLLAPSE	9	distinguished	0				sepereted	0
EARNED	9	divided	0				set	0
HOLDS	9	DO	1				showed	1
LED	9	done	1				showen	1
RELATED	9	drivided	0				shown	1
RETAIN	9	examined	0				splited	0
DID	8	exist	0				tell	0
EMPHASIZED	8	existed	0				the	0
FIND	8	explain	0				told	0
KEEP	8	explained	0				understand	1
MAY	8	fascinating	0				understood	1
SUPPORT	8	finded	1				use	1
WERE	8	finished	0				used	1
ILLUSTRATE	7	found	1				varied	0
							witnessed	0

Appendix F: Item-total statistics for dichotomous and polytomous data

Item-Total Statistics Dichotomous

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	5.5049	21.49	-0.006	0.834
2	5.5146	20.758	0.305	0.827
3	5.4369	20.355	0.33	0.826
4	5.4417	20.794	0.196	0.83
5	5.568	21.29	0.189	0.83
6	5.3786	20.441	0.256	0.829
7	5.3641	19.696	0.454	0.821
8	5.5243	20.66	0.375	0.825
9	5.301	19.704	0.408	0.823
10	5.2718	20.355	0.233	0.83
11	5.534	21.177	0.156	0.83
12	5.5583	20.97	0.358	0.827
13	5.5485	21.156	0.2	0.829
14	5.5534	21.312	0.123	0.83
15	5.4223	20.362	0.313	0.826
16	5.5485	21.068	0.25	0.828
17	5.4272	19.904	0.461	0.822
18	5.3641	19.559	0.492	0.82
19	5.4417	20.375	0.329	0.826
20	5.1893	19.803	0.345	0.826
21	5.5728	21.524	0.011	0.832
22	5.267	19.523	0.438	0.822
23	5.5631	21.145	0.269	0.828
24	5.4175	19.698	0.512	0.82
25	5.2184	19.869	0.336	0.826
26	5.1553	19.498	0.412	0.823
27	5.5631	21.252	0.193	0.829
28	5.5777	21.416	0.138	0.83
29	5.4272	20.158	0.381	0.824
30	5.199	18.804	0.592	0.815
31	5.4126	20.302	0.322	0.826
32	5.4612	20.191	0.417	0.823
33	5.4466	20.317	0.354	0.825
34	5.5485	21.078	0.245	0.828
35	5.2476	19.338	0.476	0.82

Item-Total Statistics

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	22.6505	121.985	0.029	0.893
2	22.9223	119.643	0.244	0.89
3	22.3301	116.769	0.371	0.888
4	22.3883	118.668	0.233	0.89
5	22.5825	119.132	0.288	0.889
6	22.3883	117.224	0.289	0.89
7	22.5	115.285	0.383	0.888
8	22.3641	117.696	0.388	0.888
9	22.0631	115.055	0.453	0.886
10	22.0534	115.163	0.429	0.887
11	22.2767	118.855	0.329	0.888
12	22.9563	121.466	0.124	0.891
13	22.4951	117.227	0.431	0.887
14	22.8058	118.333	0.354	0.888
15	22.6019	116.055	0.375	0.888
16	22.5728	116.587	0.476	0.886
17	22.4417	114.219	0.508	0.885
18	22.267	113.465	0.535	0.885
19	22.3107	113.962	0.578	0.884
20	21.8592	115.009	0.479	0.886
21	22.335	117.902	0.47	0.887
22	22.3252	112.357	0.488	0.886
23	22.8544	119.92	0.251	0.889
24	22.4709	113.041	0.569	0.884
25	22.0583	113.401	0.488	0.886
26	22.0146	111.644	0.557	0.884
27	22.6311	115.941	0.558	0.885
28	22.6942	118.652	0.35	0.888
29	22.6311	116.107	0.376	0.888
30	22.1165	109.86	0.648	0.882
31	22.3592	114.826	0.47	0.886
32	22.6214	114.939	0.479	0.886
33	22.4806	115.714	0.424	0.887
34	22.8252	119.794	0.238	0.89
35	22.0534	113.465	0.512	0.885